Embracing nature's inhomogeneity –

the challenge to infer spatio-temporal dependences from paleoclimate data

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
Dr. rer. nat.
im Fach Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Dipl.-Phys. Kira Rehfeld

Präsident der Humboldt-Universität zu Berlin:
Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Stefan Hecht, PhD

Gutachter:
1. Prof. Dr. Dr. h.c. mult. Jürgen Kurths
2. apl. Prof. Dr. Holger Lange
3. Prof. Dr. Igor Sokolov

eingereicht am: 8.1.2013
Tag der mündlichen Prüfung: 30.5.2013

"The butterfly counts not months but moments,
and has time enough."
Rabindranath Tagore, Fireflies

Abstract

Investigating past climate changes offers a unique key to understanding the future behavior of the Earth system under anthropogenic perturbation, because it is the only realization of the "Earth system experiment" accessible. Paleoclimate archives such as trees, stalagmites, or glacial deposits provide in their structure and composition time-dependent records of earlier climatic variability. Statistical analysis of dependencies amongst such time series helps to infer on the climatic processes reflected in the paleoclimate proxy data and then, ultimately, on the dynamics of the Earth system. Three inherent technical challenges need to be met: the datasets are heterogeneously sampled in (i) time and (ii) space, and time itself is a variable that needs to be reconstructed which (iii) introduces additional uncertainties.

To address these issues I developed the paleoclimate network framework, inspired by the increasing application of complex networks methodology in climate. I introduced estimators for Pearson correlation, mutual information and event synchronization that do not require time series sampled at regular intervals. The impacts of age uncertainty on such similarity estimates was assessed numerically, using ensembles of possible accumulation histories. A simple model for information flow in the Asian summer monsoon (ASM) was used to test the ability of (paleoclimate) network measures to detect spatio-temporal transitions from time series observed at heterogeneous locations in space. The Gaussian-kernel based estimators for Pearson correlation and mutual information were more efficient for irregular time series than standard estimators applied to interpolated time series. The proposed event synchronization function quantifies similarity between time series based on the relative timing of extreme events in the time series. In benchmark tests I found it especially suitable for irregular and age uncertain time series. In principle, also heterogeneously sampling in space did not preclude the detection of spatio-temporal transitions by the proposed paleoclimate network measures. Using ensembles of model realizations I found that measures such as the proposed cross-link ratio and regional node strength reflect these changes consistently, both when estimated from a grid-based realization and when the model was sampled at available paleoclimate archive locations. In contrast to this, shortest path betweenness, a popular measure for complex networks, did not reflect these transitions. I applied the paleoclimate network approach to investigate decadal scale dynamics in the Asian summer monsoon system for the last millennium. Specifically, I tested to what extent a possible temperature-induced spatio-temporal change in internal ASM dynamics could be discernible given that the dataset available is age uncertain, sparse, and irregularly sampled in time and space. For the given dataset I found that age uncertainty and data sparsity precludes robust estimation of dependencies at decadal resolution. The distribution of link strength in the network did not depend significantly on the type of paleoclimate archive from which the records came, or on the distance between their origin. This could indicate that global-scale teleconnections, rather than local convective dynamics, are reflected in these paleoclimate records.

The presented paleoclimate network approach is suitable to integrate methods that address the challenges in the reconstruction of paleoclimate dynamics. Future improvements could be sought by integrating directed measures of statistical association, for example to investigate the direction and strength of a potential coupling between the ASM and the El-Niño phenomenon in the past.

Zusammenfassung

Die Untersuchung vergangener Klimavariabilität ist ein einzigartiger Schlüssel zum Verständnis zukünftigen Verhaltens des Erdsystems unter anthropogener Einwirkung. Dies ist von besonderer Wichtigkeit, da es die einzige Realisierung des „Erdsystemexperiments“ ist, die für uns zugänglich ist. Paleoklimaarchive, wie Bäume, Stalagmiten oder Gletscher stellen in ihrer Struktur und Zusammensetzung zeitabhängige Aufzeichnungen früherer Klimavariabilität dar. Die statistische Analyse von Zusammenhängen zwischen solchen Zeitreihen kann helfen, die den Paläoklimaproxies zugrundeliegenden Klimaprozesse und, letztlich, der Erdsystemdynamik zu verstehen. Drei Hauptherausforderungen müssen gemeistert werden, um dies möglich zu machen: die Zeitreihen sind unregelmäßig aufgelöst in (i) Zeit, (ii) Raum und die Zeit selbst ist eine Variable die rekonstruiert werden muss, was (iii) zusätzliche Unsicherheiten mit sich bringt.

Dazu habe ich den Paläoklimanetzwerkansatz entwickelt, inspiriert von der zunehmenden Anwendung von Methoden aus dem Bereich der komplexen Netzwerke in der Klimatologie. Ich habe Schätzer für Pearson-Korrelation, Transinformation (Mutual Information) und Ereignissynchronisation (Event Synchronization) eingeführt, die keine Zeitreihen mit regelmäßigen Beobachtungsintervallen benötigen. Der Einfluß von Altersunsicherheiten auf Schätzungen solcher Ähnlichkeitsmaße wird numerisch durch Ensembles von möglichen Akkumulationsverläufen abgeschätzt. Ein einfaches Modell für Informationsflüsse im Asiatischen Sommermonsun (ASM) ermöglicht den Test der Fähigkeiten von (Paläoklima-)Netzwerkmaßen, räumlich-zeitliche Klimaänderungen von Zeitreihen räumlich heterogen verteilter Orte zu detektieren. Die Gauß-Kernel-basierten Schätzer für Pearson-Korrelation und Transinformation sind effizienter für unregelmäßig abgetastete Zeitreihen, als die Standardschätzer, wenn sie auf interpolierte Zeitreihen angewendet werden. Die vorgeschlagene Ereignissynchronationsfunktion quantifiziert Ähnlichkeiten zwischen Zeitreihen basierend auf dem relativen Timing von Extremereignissen in den Zeitreihen. In Benchmark-Tests stellte ich fest, dass die Funktion besonders für unregelmäßige und zeit-unsichere Zeitreihen geeignet ist. Im Prinzip schließt auch die unregelmäßige Abtastung von Klimaprozessen im Raum die Detektierbarkeit räumlich-zeitlicher Veränderungen durch die vorgestellten Paläoklimanetzwerkmaße nicht aus. Mittels Ensembles von Modellrealisierungen fand ich, dass Maße wie die vorgeschlagene *Cross-Link-Ratio* und *regionale Knotenstärke* diese Veränderungen konsistent abbilden, sowohl aus gitter-basierten Realisierungen wie auch wenn das Modell an Orten, an denen Paläoklimadaten zur Verfügung stehen, beprobt wurde. Im Gegensatz dazu spiegelte die für komplexe Netzwerke populäre *Shortest Path Betweenness* die Veränderungen in keinem Fall wieder. Ich habe den Paläoklimanetzwerkansatz angewandt, um dekadische Klimadynamik im ASM für das vergangene Jahrtausend zu untersuchen. Konkret testete ich, inwieweit eine mögliche temperatur-induzierte Klimaveränderung in interner ASM-Dynamik unterscheidbar ist, wenn die Altersunsicherheiten, die niedrige Anzahl und die unregelmäßige Auflösung in Raum und Zeit berücksichtigt werden. Für den zur Verfügung stehenden Datensatz verhinderten Altersunsicherheiten und die niedrige Zahl der Daten eine robuste Schätzung von Abhängigkeiten auf dekadischer Zeitskala. Ein Zusammenhang zwischen der Art der Paläoklimaarchive, oder der Entfernungen zwischen ihnen, und der Linkstärkenverteilung im Netzwerk fand sich nicht. Dies könnte darauf hinweisen, dass sich globale Fernwirkungen, mehr als lokale konvektionsbasierte Dynamik, in diesen Paläoklimazeitreihen spiegelt.

Der vorgestellte Paläoklimanetzwerkansatz zeigte sich geeignet, die Herausforderungen in der Rekonstruktion von Paläoklimadynamik flexibel zu addressieren. Zukünftige Verbesserungen könnten in der Integration von gerichteten Kopplungsmaßen gesucht werden, zum Beispiel um die Rekonstruktion der Richtung eines möglichen Kopplungsmechanismus zwischen dem ASM und dem El Niño-Phänomen in der Vergangenheit zu ermöglichen.

Contents

1 Introduction

Nature won't dance to humankind's tunes – it has not in the past, it has not yet in the present, and it probably will not in the future. Instead, ancient cultures have collapsed when climates became less hospitable, from the Maya in Central America [77] to the Indus Valley Civilization [149] and dynasties in China [190]. Investigating such past climate changes offers a unique key to understanding the future behavior of the Earth system under anthropogenic perturbation, because our global past is the only truthful realization of the "Earth system experiment" we are part of. The nonstationary dynamics of the climate system evolve at the interplay between its topological, chemical and thermodynamical boundary conditions and the external forcing that is exerted mainly by the sun. Yet, a closed description of the system's dynamics is impossible due to the large number of variables, continuously changing boundary conditions and nonlinear interdependencies giving rise to chaotic behavior. Climate modeling and paleoclimate reconstruction are the two central means to improve the knowledge and understanding of past Earth system dynamics. They depend on each other: Paleoclimate data are the only witnesses that can yield the knowledge about the state of the Earth system in the past; climate models try to solve fundamental equations for physical, chemical, and nowadays also biological processes, against the background of solar forcing, the Earth's rotation and global geography.

Trees, stalagmites, sediments – paleoclimate archives such as these grew in the past and accumulated information about surrounding climate processes in their structure and composition. Carefully validated, such proxy time series deliver records of ancient temperature, precipitation amount, greenhouse gas concentration, glaciation or received solar forcing. These are essential in the development and validation of climate models, whose physics should ideally be able to reproduce such climate states [148, 153, 171].

Yet, paleoclimate archives are the product of complex natural processes [180]. When they provide a suitable climate proxy, the first question to be addressed is what time period these climatic proxy observations correspond to. This time information can be gleaned from dating techniques [21, 125, 156] and by comparison to other records [111, 33], but *age uncertainty* is unavoidable, which poses the first obstacle to overcome for an understanding of the climate system. Also, the growth of the archives is itself a predominantly random stochastic process, modulated by the climate system [180] and may temporally cease altogether [104, 25, 49]. If favorable conditions prevail, proxy information might be recorded at a higher rate than if the situation was disagreeable to the tree or stalagmite, and archives are found only in suitable geographic places. This results in the next technical challenge to successful paleoclimate reconstruction: The time series of the paleoclimate proxy parameters are often sparsely and irregularly sampled over time and space, undesirable properties for statistical analysis [84]. Finally, and fundamentally, different archives and proxies might observe climate from different perspectives, due to their position in space, but also due to their individual physical archiving processes. Therefore, local climate effects need to be distinguished from global influences and archive-dependent noise.

Until now, paleoclimate research has focused predominantly on the reconstruction of climatic variables, both temporally as well as spatially. Also, rhythms inherent to the system have received much attention [12, 140, 147, 13], but the investigation of dependences within and amongst different climate subsystems and potential forcing parameters has been mostly restricted to visual inter-comparison of reconstructed proxy time series [102, 144, 190]. Considering the increasing wealth of datasets available, however, the human eye needs the assistance of statistical routines to

detect similarities amongst the records. Additionally, statistical measures of association provide a quantitative estimate of the degree of similarity, and its significance.

In this work I introduce paleoclimate networks as a framework for the systematic assessment of dependencies amongst time series of paleoclimate processes. The basic idea is inspired by the recent advances in complex networks theory and its first successful applications in climate science [165, 183, 40, 91, 150, 42, 58]. Networks are made up from *links* that connect different *nodes*. For paleoclimate networks, nodes are identified with locations, at which paleoclimate data are available. They are associated with the corresponding time series of paleoclimate proxy data. Two nodes in such a graph are linked, if their time series are statistically significantly similar. A wealth of graph-theoretical methods can then be applied to the obtained network, to assess the interdependence structure, globally for the whole network, but also locally.

The key questions that have to be addressed for successful paleoclimate dynamics reconstruction are, as illustrated in Fig. 1.1:

- How can statistical similarity be detected for extremely short and irregularly sampled time series? What systematic effects are associated with this?
- To what extent does uncertainty on the abscissa of these time series affect the estimation of statistical association?
- How are complex network statistics affected by the spatially heterogeneous distribution of the data?

Figure 1.1: The key challenges to the successful reconstruction of paleoclimate dynamics are temporal and spatial heterogeneity and the age uncertainties associated with the paleoclimate proxy records.

Similarity between time series can be estimated for example by Pearson correlation, which quantifies how well the relationship between two time series can be described by a linear function. Mutual information, a measure originating from information theory, is able to assess nonlinear dependence strength. The commonly employed algorithms to infer the extent of association from time series, however, require perfectly coinciding observation times. Interpolation methods are commonly employed to impose such conformity, but using them introduces additional, sampling dependent, errors and artifacts [140, 4, 134, 135, 38]. This is not only a problem in the geosciences and ecology [140], but also in astronomy [134, 135, 136], turbulence research [4, 64], economy [37] and biomedical applications [137, 39, 68], and some promising methods have been put forward.

Some studies found, that analyses of irregular time series is less prone to aliasing and can be taken to temporal resolutions beyond the mean data rate [22, 168], thus the irregularity is sometimes even introduced for this purpose [105, 169].

Analogously, the climate network methodology as developed up to now assumed that the nodes are distributed on a regular grid. Although spatio-temporal gap-filling algorithms exist [78] and are being applied for modern reanalysis data, there is a fundamental reason why interpolation, at least in the paleoclimate context, is prohibitive: Spatially, it is unclear, whether proxy information from close by archives can be used to infer on climate variability in a given location. Temporally, climate variability might have been fundamentally different when the archive was actively growing, than when it was not. To infer from one on the other, therefore, puts intended analyses on very shaky ground.

The test region for the paleoclimate network approach I put forward in this work is the Asian monsoon domain. Spatio-temporal transitions have occurred and are underway in many regions on Earth. However, the Asian monsoon is a climatic phenomenon with global significance, as more than 60% of the world‘s population depends on it critically [172]. So both long- and short-term variability in the Asian monsoon system are a cause of concern in all South Asia, because their effects, through teleconnections, affect climate globally [174, 167, 28]. Extremes in the dynamics of the monsoon systems can lead to droughts and floods and affect both the people as well as the economies of the region [90, 91, 82]. The interaction amongst the different subcomponents of the Asian monsoon, as well as its drivers, remain far from being fully understood [174, 177]. In this context I develop a simple model for information flow through convective and diffusive fluxes in the Asian summer monsoon. I use it to study whether and how spatio-temporal transitions are reflected in (paleo-)climate network measures, and to investigate how heterogeneous spatial sampling affects such inference. The model can be used to generate *pseudo-proxies*, time series that mimic spatial and temporal sampling of real archives. While previous pseudo-proxy studies have been based on computationally heavy global circulation models [98, 171, 148, 83], KIMONO is far less demanding and can therefore be used for ensemble studies. I apply the developed paleoclimate network approach, and the semi-empirical model KIMONO to test for a potential temperature dependence of the Indian summer monsoon influence in the Asian Monsoon domain.

To summarize, the key idea, along which I want to address the aforementioned challenges, is to develop methods for the extraction of spatio-temporal dynamics that do *not* require re-sampling of the available paleoclimate data onto a uniform scale, neither in time nor in space. Using numerical simulation and benchmark tests I aim to infer the relevance of sampling irregularity, sparsity and time scale uncertainty.

Paleoclimate reconstruction poses difficult challenges to statistical inference. I address some of them in the course of this thesis, which is outlined below:

Chapter 2 targets the basic questions revolving around similarity assessment for irregularly sampled time series, and develops adapted kernel-based estimators and a benchmark test for synthetic data to assess the performance of similarity estimators under heterogeneous temporal sampling.

Chapter 3 focuses on the effects of age uncertainty on similarity estimation for time series using numerically simulated proxy data and tests, which similarity estimators from Chapter 2 perform best for age uncertain, autocorrelated and short time series.

Chapter 4 puts forward the paleoclimate network approach and investigates the effect of spatial heterogeneity using the semi-empirical model *KIMONO* which models information flow in the Asian monsoon domain due to convective phenomena.

Chapter 5 applies the developed paleoclimate network approach to investigate Asian summer monsoon dynamics for the past millennium on short, decadal, time scales. In combina-

tion with the *KIMONO* model I assess data quality limitations posed to the evolutionary reconstruction of paleoclimate dynamics.

Chapter 6 summarizes and interprets the main findings and outlines of future work.

Appendix 1 describes the derivation of the variance factors for the KIMONO model, and presents supplementary results for the ASM paleoclimate network.

Some of the material presented in this thesis has been already published. Specifically the "comparison of correlation analysis techniques for irregularly sampled time series" [128] in Chapter 2, and the review of literature on the Asian monsoon [129] in Chapter 5 formed part of the publications listed on page 129.

2 Similarity measures for irregularly sampled time series

In this chapter I will first give a background on (paleo)climate time series in Sect. 2.1, relevant definitions (Sect. 2.3) and motivate why the robust and efficient assessment of similarities (Sect. 2.4) between these time series is of high importance in understanding climate dynamics. I then develop and test novel estimators for Pearson correlation (Sect. 2.5), mutual information (Sect. 2.6) and a event synchronization function (2.7) and illustrate their ability to cope with irregular, non-even observation time intervals. The notion of a *strength* of de-facto similarity, as developed in Sect. 2.8, combines the results of these estimators for their simultaneous use in the paleoclimate network framework.

2.1 Paleoclimate time series

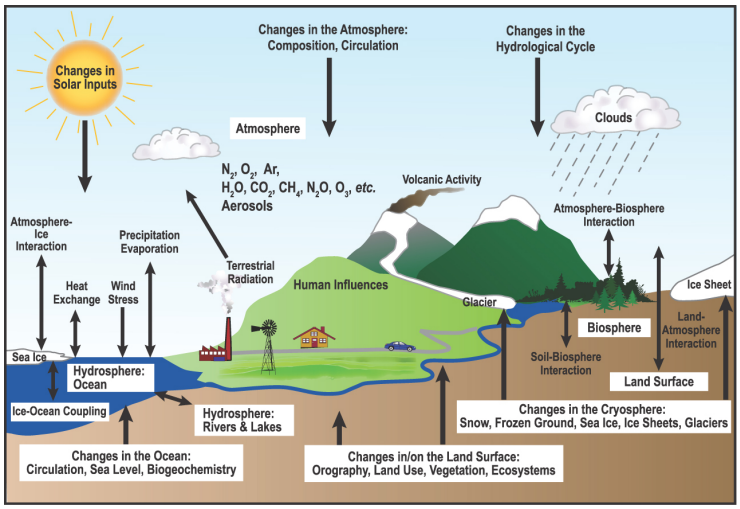


Figure 2.1: Schematic view of the climate system, its processes and interactions, adapted from [72].

The complexity of the Earth system and its large number of degrees of freedom poses a difficult challenge for geoscientists and climatologists (Fig. 2.1). While the system is, in fact, many-dimensional, only few variables can generally be observed, and this only at discrete times. Assuming that the climate system could be fully described as a (time-dependent) system of interdependent processes (e.g., within the atmosphere, the oceans, the biosphere), its *state* could be given by time-dependent equations [155]. Characterizing the state of the system, consequently, would therefore require measuring ('observing') all state variables that describe it, e.g. precipitation amount, wind speeds, temperatures, and time. Even if this was feasible, the changing topography of the Earth also plays an important role in forming climate dynamics, and thus additional components for such changes of the system's boundary conditions would need to be added to the equations describing the dynamics in the system. On longer time scales, even if all parameters in the governing equations could be identified, a closed solution and prediction of this

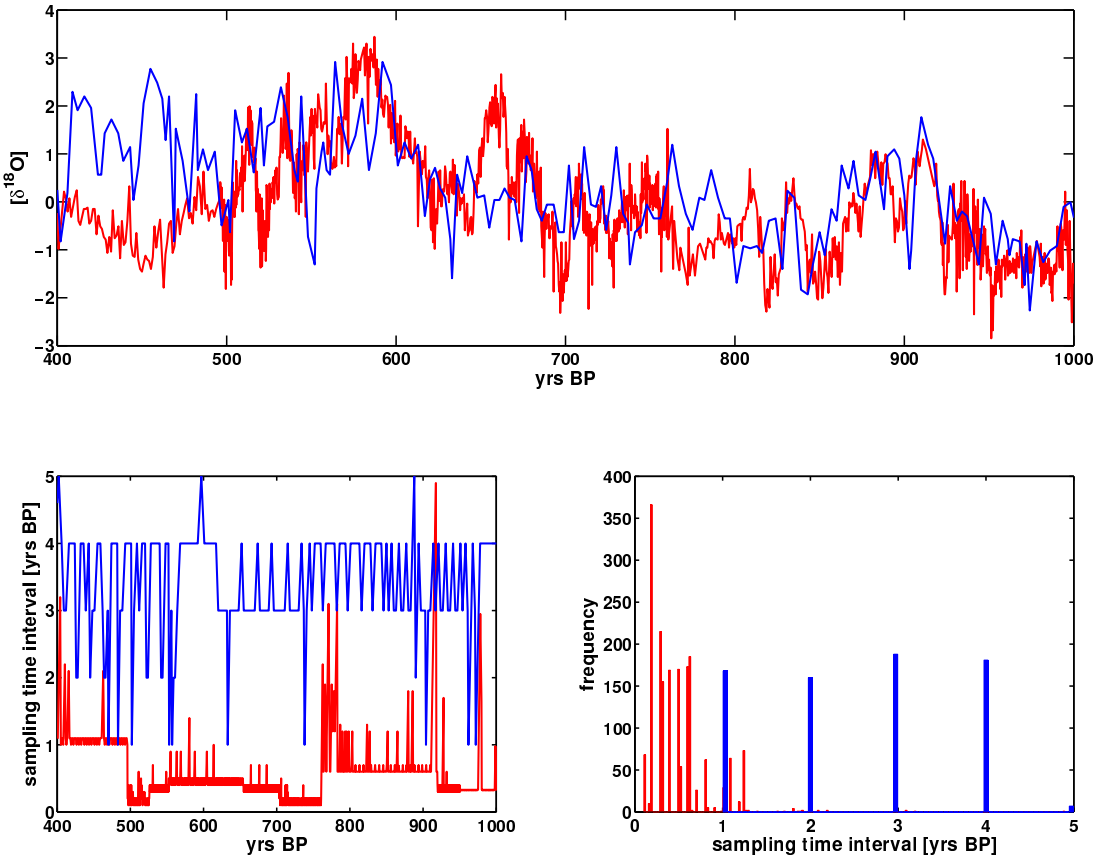


Figure 2.2: Top: Paleoclimate time series for Dandak [13] (blue curve), and Wanxiang [190] (red curve) caves, situated in India and North-East China, respectively. The sampling, or reconstruction, rate varies over time (bottom left) and the inter-sampling time distributions (bottom right) are different. While there are times when the curves show good agreement, visually, at other times excursions are not matching, which could indicate the influence of local, non-climatic, noise, or a change in the processes dominating both climate regimes.

nonstationary and high-dimensional system would remain impossible due to nonlinearities and stochasticity. What remains is the challenge to infer as much information on the system's dynamics as possible. Two approaches are combined to improve our understanding of the earth system: climate modeling, and paleoclimate data analysis [66, 171]. Climate models solve conservation equations and integrate knowledge on physical mechanisms, and paleoclimate reconstructions provide boundary conditions and, ideally, ground truth to test model results.

In paleoclimatology, target parameters may include for example annual temperature, precipitation amounts, productivity in the oceans or in a lake or average wind speed. These can be computed from variations in *paleoclimate proxies*, measured in *paleoclimate archives* such as old trees, ice cores, stalagmites or sediments below water bodies [180]. Proxies are named either after the structural or compositional property of the archive which they describe, or after the climatic parameter they are calibrated for [180]. In principle, proxy variability P is a function of climate and recording noise ε

$$P = \mathcal{F}(\text{climate}) + \varepsilon \tag{2.1}$$

which, via *proxy calibration* is narrowed down to the dependency of P on climatic parameters $\mathcal{C}$

such as, e.g., temperature, precipitation amount, or average wind speed.

$$P \approx \mathcal{F}(\mathcal{C}) + \varepsilon \ . \tag{2.2}$$

Although F is often assumed as being a linear function of a single climatic parameter [180, 171], it actually often is not [95, 59], which underlines the necessity to use nonlinear similarity measures in the analysis of paleoclimate proxy time series.

Temporally, these archives *sample* climate dynamics at different *accumulation rates* of the archive material. The time resolution of the measured proxy depends on the time scale of the sampling, and is often highly irregular. For example, stable isotope ratios [1] may be proxies for ambient atmospheric temperature or precipitation. An example of two time series, where $\delta^{18}O$ is interpreted as indicating monsoonal precipitation amounts, related to Indian Summer monsoon strength, is given in the top panel of Fig. 2.2. Similarly, annual tree ring width may be related to the length of the growing season or the amount of growth-limiting irradiation or precipitation [71].

At each position $\vec{p_i}$ on Earth, local climate can be described by the multivariate set of observations $\vec{X}_t(\vec{p_i})$, containing all necessary variables to describe the (thermodynamic) state at this point in space [57]. This local climate process is a function of all other local climate processes $\vec{X}_t(\vec{p_j})$ in all locations $j = 1, \ldots, N$ on Earth that represent the internal to the Earth system, potential external forcing $\vec{F}_t^{\text{ext}}$ and time:

$$\vec{X}_{t_0}(\vec{p_i}) = f\left(\sum_j^N \vec{X}_{t \leqslant t_0}(\vec{p_j}), \vec{F}^{\text{ext}}_{t \leqslant t_0}, t \leqslant t_0\right) \ . \tag{2.3}$$

The dependencies of local and global climate dynamics on each other, and on external forcing, are highly complex and nonlinear [73, 12]. Different components of the system evolve on different timescales, therefore climate states at one place $\vec{p_i}$ and time t_0 depend causally on the states in all places $\vec{p_j}$ at all times prior and equal to t_0, $t \leqslant t_0$. In principle, and presuming the coupling structure between the components are known, Equ. 2.3 could be formulated as a system of differential equations, combining the Navier-Stokes-Equations [153, 57], describing fluid motion under a conservation of mass, momentum and energy, with thermodynamic interactions between solar radiation, ocean and atmosphere. External forcing is also exerted by gravitation, and Coriolis forces generated by the Earth's rotation.
But we are far from being omniscient in this regard. Climate models integrate such equations for a reduced set of variables and forcing factors. Their development requires the identification of a) relevant climate processes, including dependencies and time delays and b) assumptions on the forcing structure, extent and sensitivity of the system and c) knowledge of the boundary conditions given by topography and geology. Paleoclimate data provide information on boundary conditions (e.g. geochemistry, ice volume, geomorphology) and forcing (irradiation)[171, 153]. Reconstructions of temperature [186, 96, 98, 158], precipitation and vegetation provide limited, but crucial, data on the underlying dynamics.

To understand components of the global climate system, not only from the paleoclimate data perspective, it is helpful to linearize Equ. 2.3. It can be simplified, assuming separability of local, global and external factors and stationarity of boundary conditions:

$$\vec{X}_{t_0}(\vec{p_i}) \approx f_1(\vec{X}^{\text{loc}}_{t_1 \ldots t_0}) + f_2(\vec{X}^{\text{glob}}_{t_2 \ldots t_0}) + f_3(\vec{F}^{\text{ext}}_{t_3 \ldots t_0}) \ . \tag{2.4}$$

In this form, the climatic process $\vec{X}_{t_0}$ at point $\vec{p_i}$ is given by the linear combination of surrounding

[1] $\delta^{18}O$ refers to the relative number of oxygen atoms with an atomic weight of 18 vs. the number of atoms which are lighter, with an atomic weight of 16 [for more details see, e.g., reference 180].

local climatic processes f_1, processes on the global scale f_2 and external forcing f_3 on the relevant time scales t_1, t_2 and t_3. Cross terms, such as for example the dependency of global climate (f_2) on external forcing (f_3) are neglected. Still, this to some extent phenomenological and crude equation is useful to understand, what insights paleoclimatic time series can provide: The spatio-temporal coherence of local climatic processes can be assessed by comparing paleoclimate proxy time series stemming from one region. The dependency of regional climate on global changes can be gleaned by investigating similarities between time series from one region (e.g. rainfall reconstructions from South Asia) and proxy reconstructions of processes that influence climate on a global scale (e.g. polar glaciation). On one hand, these relationships are modeled using coupled global circulation models, which solve continuity equations for basic physical properties at each point in space and time, stimulated by prescribed or generated external forcing. On the other hand, basic hypotheses on how the Earth system works are generated by investigating the interdependencies in past and present climate data and can later be tested using climate models.

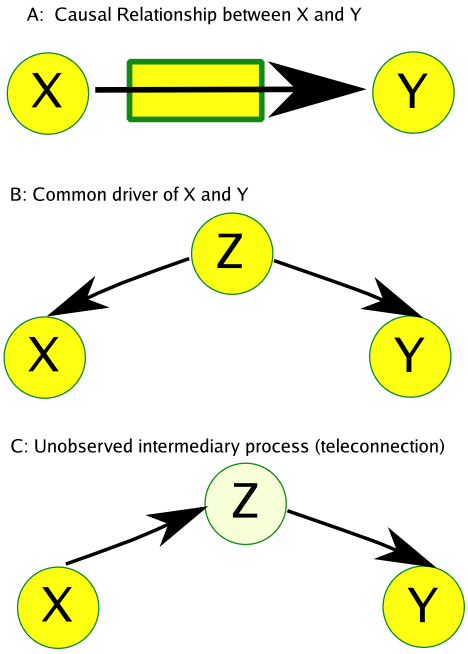


Figure 2.3: Schematic overview of possible causes for actual interdependencies between two processes X and Y: Causality, with X determining Y (A), common variability due to a common driver (B) and influence through an intermediary process (C).

Having observed two climate processes at different points on Earth, the similarities between the time series may, as illustrated schematically in Fig. 2.3, indicate

- that there was *flow* of matter, energy or momentum between the locations, thereby synchronizing local climates (Fig. 2.3A) or
- that climate dynamics at both locations was influenced by a *common driver*, or external forcing (Fig. 2.3B),
- that there was a teleconnection, caused by intermittent, unrecognized, processes (Fig. 2.3C).

Unfortunately, our common understanding of the climate system is hampered by the fact that only *one* realization of the stochastic Earth system 'experiment' is under way, and for the most

part of the Earth's history, climate parameters were not accurately observed. Therefore proxy reconstructions for past climate parameters, and the imposed forcing, are crucial to understanding the climate system, and its potential reactions to anthropogenic perturbations. In this context, the spatio-temporal significance of reconstructed paleoclimate time series, and the dependency on global or external forcing, is often inferred from similarities between graphical visualizations of the time series (coinciding maxima/minima), and sometimes the computation of correlation statistics.

A crucial problem with these records it their irregular sampling in time due to the complex sedimentation/ accumulation rate. Standard methods can not be applied when timescales and resolutions are different. This is not a problem in the geosciences only, as irregular observation of continuous-time processes also occurs in the detection of biomedical rhythms [137], astronomy [45, 134, 135, 136] or turbulence research, where the velocity of the flow can only be measured if seeding particles pass a measurement volume [23, 64].

2.2 Interpolation of irregularly observed paleoclimate time series

Paleoclimate archives sample local climate variability as they grow. As this growth is a natural process this results in irregular *sampling times* for the proxy reconstruction. When the analysis of paleoclimate records is taken beyond visual comparison, the time series are usually interpolated to a common time scale so that classical analyses requiring bivariate observations can be performed. For examples see for example [190, 33, 102].

Interpolation techniques are understood, in the context of this work, as procedures that *resample* the originally irregular vector of observation times (t_x) of length N_x with a 'non-delta-like' distribution of the inter-sampling time observations $(\Delta t_x) = (t_x(i) - t_x(i+1))_{i=1\ldots,N_x-1}$. This distribution of observation time distances can be characterized, for example, by its *mean*, *variance* and its *skewness*. Two examples for actual paleoclimate records are given in Fig. 2.2: While the chronology for the Dandak record has been reconstructed such that it has regular observation intervals, but is *missing values* for one, two or four consecutive years, the Wanxiang record provides more information on sub-annual timescales.

To obtain coeval observations for both records by interpolation, an interpolation step Δt_r, tied to the mean sampling intervals of the records Δt_x, is usually chosen as $\Delta t_r = \max(\Delta t_x, \Delta t_y)$. The available proxy observations are then interpolated to the resampled time vector $\Delta t_r = (t_r(i) - t_r(i+1))_{i=1\ldots,N_r-1}$, which has equally spaced inter-observation times t_r, as $\Delta t_r = \frac{t_r(N_r)-t_r(1)}{N_r}$. In practice, there is no convention on the resampling rate in the Geosciences [140, 108].

Any form of interpolation makes an implicit assumption on the development of the processes observable. For the commonly chosen linear interpolation technique, the reconstructed observation $x_r(i)$ at the time point $t_r(i)$ lies on the straight line connecting the closest preceding observation $x(j)$ at time point $t_x(j)$ with the observation $x(j+1)$ at a later time point $t_x(j+1)$. The falseness of such a simple assumption on proxy behavior becomes apparent – and especially relevant – when longer gaps in the time series exist. It has been shown that applying standard interpolation techniques, also beyond simple linear forms, to originally irregular time series causes positive spectral bias towards low frequencies and consequently high-frequency variability is underestimated [4, 128, 140, 154]. This particular effect could be overcome by resampling – or gap-filling – algorithms that preserve high frequency variability, using Singular Spectrum Analysis [78] or Lomb-Scargle periodogram inversion [67]. The conjectured sinusoidal periodicity upon which these methods are based may be more or less valid, but still the fundamental assumption remains the same: That the climate process at the non-observed states resembles that during the observable periods. This is violated for example for stalagmites, which rely on a steady supply of drip water to the cave environment. Surrounding climate processes during

archive growth (e.g. with sufficient moisture availability) and impeded growth (e.g. in a drought period) are recorded at high versus low, or no, time resolution and are potentially very different. Inferring from observations of one on potential observations of the other ignores this difference.

The spectral effect of interpolation can be understood if one considers the available time scales employed for spectral analysis. All possible observation time distances in one of the time series are given by $(t_x(i) - t_x(j))_{j \neq i, j=1\ldots,N_x}$. For regular sampling $N_r - 1$ observations are available on the time scale of $1 \cdot \Delta t_r$, and for a lag time scale of $k \cdot \Delta t_r$ $N_r - k$ observations, thus decreasing linearly. As illustrated in Fig. 2.4, this is not the case for irregular time series of the same mean sampling rate: More observations are found at shorter time scales and fewer at longer distances. Interpolation to the mean sampling rate thus results in a loss of information in high frequency components, and an overestimation of low frequencies. If the statistical procedures for the analysis of such records can be adapted to ignore the sampling heterogeneity such bias effects can be avoided. Therefore, the central idea of this chapter is to develop similarity estimators that abstain from from signal reconstruction. They are tested for their suitability tested for suitability for paleoclimatic reconstruction in Chapter 3.

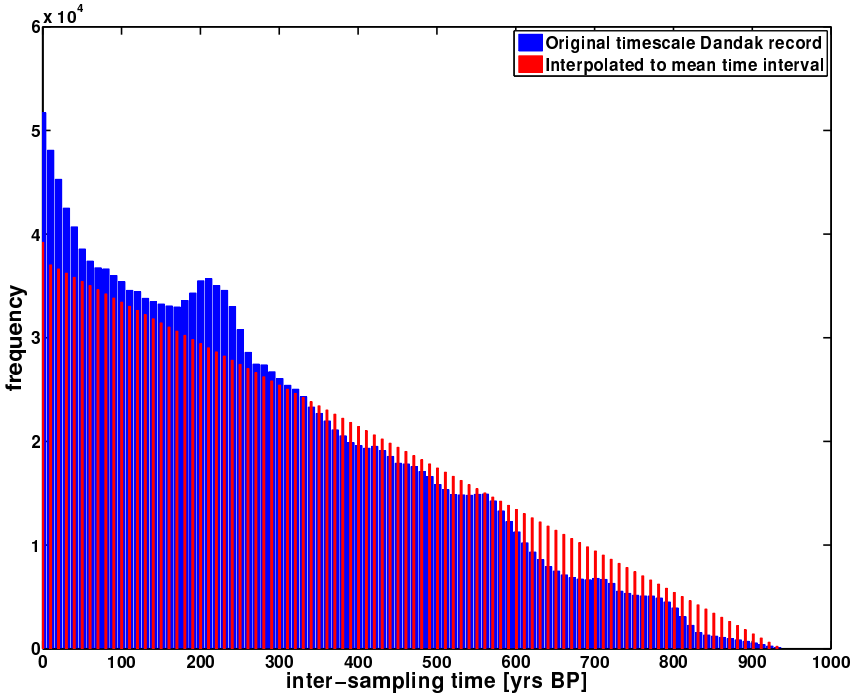


Figure 2.4: Illustration of the effect of interpolation to a mean sampling rate on the example of the Dandak cave record: The number of possible observation time distances $|t_x(j) - t_x(i)|$ in the time series depends on the chosen time scale. For a regularly sampled time series, this number decreases linearly. For an irregular time series of the same mean sampling rate *more* information is available on short timescales, and *less* on longer timescales. Naïve transformations of the originally irregular time series to a regular observation grid results in information loss because it over-emphasizes longer, and under-estimates shorter timescales.

2.3 Relevant time series definitions

In classical time series analysis the observation times are expected to be regular and certain, and the observation values to be measured exactly, as well.

Definition 1 (Regular time series) *The process X_t was observed through the regularly sampled time series $X = (x_i)_{i=1,\dots,N}$, where the observation times are given by multiplication of the index variable with the common time step: $t_i = i\Delta_t + t_0$.*

In contrast to this, for irregular time series no unique sampling rate can be defined, and the observation times cannot be directly related to an index anymore, but have to be given explicitly for each measurement. The observation time of irregular time series might even carry some amount of information independently of the observation values. drought

Definition 2 (Irregular time series) *An irregular time series $X = (t_i, x_i)$ is defined by its observation times t_i and the respective observations x_i, where $i = 1, \dots, N$.*

Input data to age modeling are (i) a dating table with its entries containing depths, associated age estimates and their uncertainties, usually given as standard deviations, and (ii) the proxy observations.

Definition 3 (Dating table) *A dating table $\mathbb{D} = (D_i, T_i, \sigma_{T_i}, \sigma_{D_i})_{i=1,\dots,N_{dat}}$ contains N_{dat} pointwise age estimates T_i taken at depths D_i, their corresponding age standard deviations σ_{T_i}, and the sample size (in depth direction) of δ_{D_i}.*

Definition 4 (Proxy observation series) *Proxy observation series $X^d = (d_j, x_j)$ are given for $j = 1 :, \dots, N_{obs}$ measurement depths d_j and proxy measurements x_j.*

Definition 5 (Age model) *For paleoclimate archives, the ages at few depths are estimated, with some uncertainty. Age models are then created to interpolate from these few dates to a time axis for the proxy time series, which is sampled much more densely in depth than the dating table. Thus, an age model is defined here as one potential depth-age relationship $t_i(z_i)$ out of the possible ensemble of age models $\mathbb{T}$.*

For Monte Carlo (MC) age modeling, whole *ensembles* of age models, $\mathbb{T}$ are created, sampling the probability space inherent in the dating table (cf. def. 3). By convention, usually the *most likely* age model is selected as the time axis for proxy time series [21, 138].

The dating table is combined with the proxy observation series using a single age model and forms a time-uncertain time series.

Definition 6 (Time-uncertain time series) *Time-uncertain time series (t_i, x_i, σ_{t_i}) assemble $i = 1, \dots, N$ observations x_i, reconstructed (most probable) observation times t_i from $\mathbb{T}$ and the observation time uncertainty σ_{t_i}.*

If the error on the time axis can be transformed into the equivalent error on the measurement axis, e.g. using Monte Carlo techniques [21, 138, 139].

Definition 7 (Time series with measurement uncertainty) *A time series with uncertainty in the measurements (e.g., the paleoclimate proxy) is given by the set (t_i, x_i, σ_{x_i}), with $i = 1, \dots, N$ observations and measurement uncertainty σ_{x_i}.*

Definition 8 (Certain, irregular and uncertain observation times) *For regular time series, the time difference coeval observations in X and Y have to exist: $t_i^x - t_i^y = 0$. In irregular time series, the observation time distance can be nonzero, but is constant $t_i^x - t_i^y = const$. For age uncertain time series $X = (t^x, x, \sigma_{t^x})$, $Y = (t^y, y, \sigma_{t^y})$ the observation time distance is a function of the age uncertainties: $\mathbb{T}_i^{x,k} - \mathbb{T}_i^{y,l} = \mathcal{F}\left(\mathbb{T}^x(\sigma_{t^x}), \mathbb{T}^y(\sigma_{t^y})\right)$.*

For the sake of simplicity the depths D and d are expected to be monotonously increasing.

Definition 9 (Monotonicity of the age depth relationship) *Monotonicity is also an important aspect in age modeling, in that the stratigraphic sequence is expected to be monotonously increasing in age, also. This restriction is carried on into age modeling by imposing a monotonicity constraint on permitted age-depth relationships $g = (t_j, d_j)$:*

$$d_j < d_{j+1} \overset{!}{\Leftrightarrow} t_j \leqslant t_{j+1} \ . \tag{2.5}$$

Sediment towards the top, at lower depth, is expected to be younger or of the same age compared to sediment found deeper, towards the bottom [2].

2.4 Similarity concepts and measures

2.4.1 Notion of similarity

Similar objects agree in some properties, while they may disagree in others. The popular saying "you can not compare apples and oranges" [3] is misleading in this context, because it implies only that it is not possible to compare them because they are not the exact *same* type of objects. It is possible, however, to compare apples and oranges in terms of their weights, glucose content, or environmental footprint, in which they might be *similar* or *different*. Similarity measures, analogously, reflect statistical properties of time series, which might have been observed from entirely different processes. Different estimators focus on different characteristic properties related to the distributions of the observations.

Time series are a collection of measurements of specific properties of an dynamical process, together with the time when the observation (or measurement) took place. The individual data points of the series are often regarded as observations of processes, which may be deterministic, stochastic, or a combination of both. For example, economical interests motivate humans to record, for example, annual wheat prices (as early as 1810AD), daily stock indices or air temperature [29]. Stock prices in different markets may be co-dependent, but the reason for similar patterns may not be the same. The processes X and Y, observed over time t, yield the time series $x(t)$ and $y(t)$. These processes, and the time series, are similar if, for example, coeval minima or maxima were observed. Comparison can then give information about functional relationships between processes underlying time series: Given that two processes X and Y are not independent, there may either be a causal relationship or they are both driven by a global *common driver*, or there are unobservable intermediate processes, as illustrated in Fig. 2.3.

If a transfer function between the two processes exists in a form of $Y_t = \mathcal{F}(X_{t+\ell})$, this results in a repetition of a pattern, though maybe distorted, that occurs in X_t at t_0 and in Y_t at a time $t = t_0 + \ell$ later. A similarity estimator quantifies the similarities in the contemporary evolution of two time series:

Definition 10 (Similarity estimator) *A similarity estimator $S = \mathcal{F}\left((t^x, x)(t^y, y)\right)$ reflects the similarity between $x(t)$ and $y(t)$ to a numeric value in an interval $[a, b]$, $S : x(t) \times y(t) \to [a, b]$.*

For most similarity measures $a = -1, b = 1$ is considered, but for different estimators different bounds exist. Here I only require that the relationship between true dependency and estimated similarity is strictly monotonously increasing. If the delay time ℓ in the transfer function is nonzero, a similarity function gives the similarity between two time series for increasing delay:

[2] Exception: *bioturbation* due to living organisms in the topmost layer of the ocean floor, or folding of geographic stratae due to tectonics [180]

[3] Or: Apples and pears, in German.

Definition 11 (Similarity function) *A similarity function $S(\ell)$ gives the estimated similarity over different lag times ℓ:*

$$S(\ell) = S(\ell \cdot \Delta t) = f\left((t^x, x), (t^y + \ell \cdot \Delta t, y)\right) \tag{2.6}$$

The spacing of the lag vector is uniform and depends on the mean time resolution of the time series: $\Delta\ell = \max(\Delta t_x, \Delta t_y)$. To indicate that I am focusing on bivariate similarity I also use the alternative notation $S(X, Y)$ which does not explicitly refer to the possible lags.

Similarity measures as required in this context should satisfy at least four properties in an adaptation of the axiomatic definition of [9]:

Symmetry: $S(x(t), y(t)) = S(y(t), x(t))$
The statistical association should not change under a permutation of the arguments.

Reflexivity: $S(x(t), x(t)) = b$
When comparing a time series with itself the dependency is always maximal.

Translation invariance: $S(x(t) + c, y(t)) = S(x(t), y(t))$, $c \neq 0$
Adding or subtracting a constant to one of the time series does not change the resulting estimate.

Scale invariance: $S(ax(t), y(t)) = S(x(t), y(t))$, $a \geq 0$
Multiplying one or both observation vectors with a constant shall not alter the estimated association.

2.4.2 Similarity estimators

Statistical dependency between time series is often estimated using *correlation* techniques, such as Pearson cross-correlation (denoted XCF in the following). XCF, which is at the heart of Sect. 2.5, is, in principle, computed by taking the arithmetic mean over the products of coeval, centralized, and standardized observations. The underlying processes are expected to be Gaussian-distributed and stationary. Each product $p(t) = x(t)y(t)$ is, in the estimation process, independent of each other and their *order* is irrelevant to the outcome of the estimation. In contrast to this, in rank correlation, such as Spearman's ρ, the individual ordering of values within each time series is compared, which makes it more flexible with regard to the required functional form of the relationship [29].

Nevertheless, nonlinear dependencies between the processes X and Y will, in general, not be correctly estimated by linear techniques. A possible alternative is to estimate the *mutual information* (MI) between the time series $x(t)$ and $y(t)$, a measure coming from information theory (Sect. 2.6). In this measure, the joint and marginal distributions of X and Y are evaluated. Its advantage is that it is model-free and able to quantify nonlinear dependencies, its disadvantage is that it is symmetric $MI(-x, y) = MI(x, y)$, so less is known about the nature of a possible transfer function between X and Y, and MI is more difficult to quantify, as the quantification is easily biased by sample size and estimator used to estimate the probability density from the histogram.

Sect. 2.7 introduces the concept of *event synchronization* (ES), which is not directly based on available time series, but on the relative timing of distinguished *events* in two time series. In its original form [123, 91] it provides a measure for the strength of synchronization and for the direction of a potential coupling between the two processes generating the events. Although differently stated in the original paper proposing this method, ES does not require regular observation intervals. Therefore it seems a promising concept for irregular time series and it will be developed into an ES similarity function in Sect. 2.7.

There are other, less common measures for statistical association or dependence. In many fields, *distance measures* are more common than *similarity measures*. Dynamic time warping, for example, is a popular *distance measure* in data mining and pattern recognition [107]. Based on the distance matrix between two (time series) vectors, the algorithm aims to find the optimal path close to the diagonal line to match the two signals, and the more it deviates from the diagonal the lower is the similarity – or the greater the distance. The estimation process is, however, sensitive to noise [107].

Similarity between two time series can also be estimated based on frequency domain information, for example by using wavelet spectra [121]. Common to all the above measures is that they rely on regular observation intervals, which may be difficult to reconstruct in paleoclimate time series.

2.4.3 Assessing the robustness and efficiency of the estimators

In the Chapters 2 and 3 the approaches listed above are investigated with respect to their suitability for estimating similarity functions for geophysical time series. Such a performance of estimators can be evaluated with respect to the 'true' *expected* dependence. This can of course only be done for modeled or synthetic time series where the association strength, underlying the processes, is accessible for example in form of auto-correlation functions (ACFs) and cross-correlation functions (CCFs), exactly.

To evaluate the different estimators the *root mean square error* (RMSE) of the estimator $\hat{\theta}$ for a statistic θ is calculated. θ can be, e.g., the cross-correlation function at lag k, $\rho_{xy}(k)$. The RMSE is given by

$$RMSE(\hat{\theta}) = \sqrt{E[(\hat{\theta} - \theta)^2]} = \sqrt{var(\hat{\theta}) + bias(\hat{\theta})^2} \tag{2.7}$$

and incorporates both variance and bias of the estimator, i.e., its variability and systematic deviation from the true value. To estimate the RMSE, a large number of time series of a given signal type and sampling scheme are generated, and the 'target statistic' $\hat{\theta}$ is computed for each. The deviation between the mean of these many estimates and the 'true' function is the approximate *bias* of the estimator and together with the variance around this mean it makes up the RMSE.

To evaluate the contribution of the sampling irregularity to the estimation error, the analysis is performed for different sampling schemes, first for regular sampling and then for more and more irregular sampling. This is done by drawing inter-sampling-time intervals from a gamma-distribution $\gamma(\alpha, \beta)$:

$$\mathcal{F}(t) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} t^{(\alpha-1)} e^{-\beta t} & t > 0 \\ 0 & t \leqslant 0 \end{cases} \tag{2.8}$$

The gamma-distribution, later also referred to as Γ-distribution, is often used to model waiting times and sedimentation processes [15, 14, 109, 140]. Shape parameter α and rate parameter β are real-valued, non-negative parameters and $\Gamma(\alpha)$ is the value of the Gamma-function. The intervals are drawn from the distribution $\gamma(\alpha, \beta)$ and subsequently concatenated into a time line for which then a corresponding signal is generated. Given α and β, the mean μ of the $\Gamma(\alpha, \beta)$-distribution is given by $\mu = \alpha/\beta$, the variance by $\sigma^2 = \alpha/beta^2$ and the skewness by $sk = 2/\sqrt{\alpha}$. For low skewness the distribution is close to normal (cf. Fig.2.11b). Since the higher order moments depend only on the shape parameter α, the scale parameter β can be varied such that the mean is kept constant while skewness and variance can increase. Exploiting this situation I will characterize the sampling irregularity by the skewness parameter in the following, as the variance $\sigma^2 = (2\beta/sk)^2 = (2\mu/(sk\ \alpha))^2$ is uniquely determined in such a parameter configuration. A distribution with a skewness of 2.85 (Fig. 2.11b) results in a time series with large gaps, as

large values become more likely in more and more skewed sampling interval distributions.

To assess the adaptability and suitability of the different estimators, a number of tests are performed on artificially generated discrete signals for which the 'true' linear dependency in form of ACFs and/or CCFs of the underlying processes is known. For each signal type at first the RMSE in the case of regular sampling is estimated. Then increasingly irregular time series with Γ-distributed inter-observation times and with increasing skewness are generated. Since the time vectors are artificial, they do not need to have an actual unit, but for simplicity I will assume that time is measured in years.

2.5 Pearson correlation for irregularly sampled data

This section is based on Rehfeld, Marwan, Heitzig, and Kurths [128].

2.5.1 Introduction

When the aim is to reconstruct the linear auto- or mutual dependencies of the underlying processes from the observations, one can estimate either (cross-) power spectra or correlation functions, as both are related to each other by the Fourier transform [29]. The irregular sampling of the time series makes direct use of the standard estimation techniques of association measures impossible, as they rely on regular observation times. For (cross-) power spectral density estimation, standard linear interpolation of these irregular observations onto a regular sampling causes an additional bias towards low frequencies in power spectral density (PSD) estimation [140].

Historically, there are several approaches to overcome this problem. The concepts can be classified into four categories: a) direct transform methods, b) slotting techniques, c) model-based estimators, and d) time series reconstruction methods [23].

The Lomb-Scargle (LS) periodogram, introduced for use in astronomy [134, 135], is a well-known direct transform method that computes a least squares fit of sine curves to the data. The obtained least squares spectrum detects peaks at high frequencies but turned out to be severely biased for turbulence spectra [23] which do not possess periodic components. If the underlying assumption of least squares optimization, that the noise in the data is normally distributed, is fulfilled, then LS is equivalent to the Maximum-Likelihood estimate. Like all least squares techniques, the estimator is not robust in the presence of outliers. This is illustrated by the limitations of the method in the application to bimodal rhythms and signals with isolated outliers [137].

Standard slotting techniques determine the correlation function by binning all available products in the lag domain, so that observations only contribute to the correlation function at a lag if their observation time difference deviates less than half the lag bin width from the considered lag. This technique was proposed by Mayo [103] in 1978 and further elaborated by [45]. It has become popular in velocimetry [23] and is frequently applied in astronomy [19, 47, 115, 188]. The disadvantage of this technique is that, without post-processing, the correlation function estimates are not necessarily positive semidefinite and the spectra computed from their Fourier transform can show negative power. Stoica and Sandgren [154], therefore, proposed a weighting technique for autocorrelation estimation which weighed observations based on a sinc kernel and claimed that it yielded positive semidefinite results. In their review, Babu and Stoica [4] also showed the application of other kernels in the time domain, including Laplacian and Gaussian kernels. The distribution of sampling time errors in time series reconstruction from paleo-archives is often assumed to be Gaussian, which intuitively supports its use in time domain analysis. Mudelsee [110] proposed two techniques to estimate the correlation coefficient that he terms 'binned correlation' and 'synchrony correlation'. 'Synchrony correlation' consists of using the percentage of pairs of observations in the different time series that have the smallest measurement time difference, treat

them as if they were observed coevally and calculate the correlation coefficient. 'Binned correlation' essentially resamples the data into time bins on a regular grid that are assigned the mean values of the observations within these bins. Using these regular, reconstructed time series, the standard correlation estimator can be applied. These two techniques are not employed here because both do not utilize all available observations individually, which means loss of information. Also, since the standard estimator is used for calculation of the correlation coefficient, binning – or resampling – is problematic when data gaps are present and the correlation function is desired.

Model-based estimators fit a model to the time series, the spectra or the ACF, which requires prior knowledge about the actual process (cf. [64] and references therein), a prerequisite which can typically not be met in the geosciences due to the heterogeneity and complexity of geophysical processes.

The fourth group of estimators resamples the data (through some kind of interpolation) in order to create time series on a regularly spaced grid, which then can be analyzed using the standard FFT-based estimators. The most frequently used technique in geophysical time series analysis is linear interpolation. Paleo data often has rather large data gaps and it is controversial if, when and how missing observations can be appropriately approximated. For standard interpolation (e.g. linear, akima-spline and cubic-spline) a significant reduction in variance toward the high-frequency range of the estimated power spectrum occurs in the analysis of irregularly sampled data [140]. When one is interested in phenomena on short timescales (compared to the mean sampling interval), such effects should be considered, and if possible, avoided.

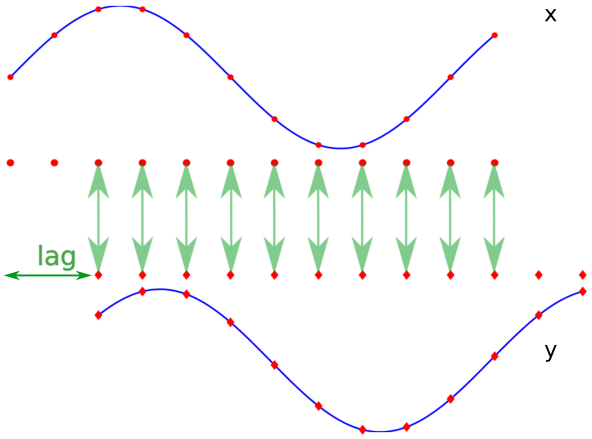
Without objective performance tests of these estimators, application of specific methods is a matter of taste, but the chosen routine may not be the optimal method available. Therefore benchmark tests comparing various methods are crucial. One study, conducted for the estimation of power spectral density from flow velocimetry data in an engineering background, has been performed by [10]. The test cases exhibited flat or simple exponentially decreasing spectra or contained a single deterministic sinusoidal component. They are therefore not nearly as complex as spectra in geophysical time series analysis typically are. Furthermore, they used a Poisson sampling scheme, which is reasonable in measurements with detector dead time, but less justified for paleo records.

Here, I will first review the methods that are or could reasonably be applied in the estimation of correlation functions of geophysical time series. This encompasses the standard approach, resampling by means of (linear) interpolation followed by a FFT-based routine, the LS periodogram, the slotting technique and kernel-weighted estimators.

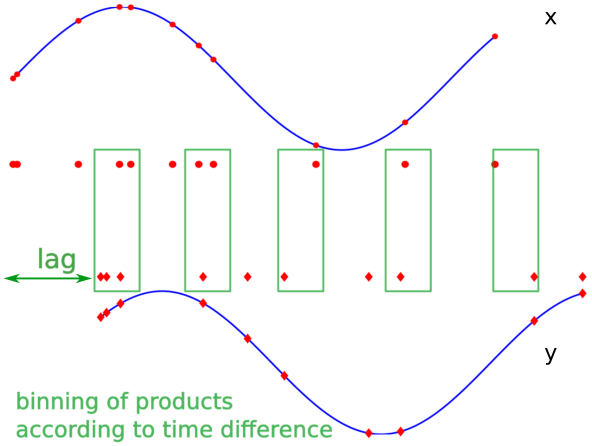
I then compare and evaluate systematically the performance of methods suitable for estimating correlation functions of geophysical time series under the presence of varying sampling schemes, and I specifically quantify the extent and direction of estimator variance and bias due to sampling irregularity. This is done using a newly developed testing scheme, based on simulated time series with increasing inter-sampling time irregularity but constant mean sampling rate.

2.5.2 Estimators for Pearson correlation

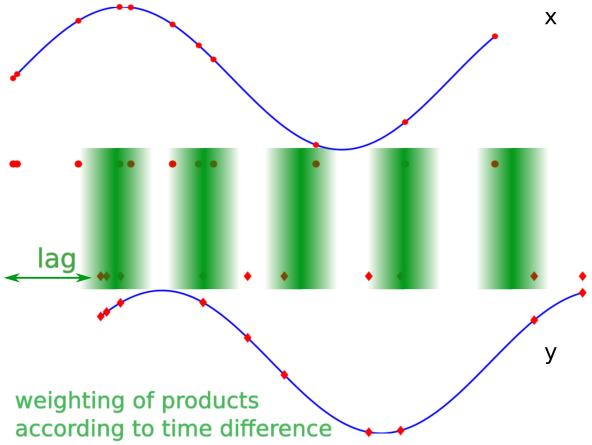
Assuming that two time series x_t and y_t were observed from stationary stochastic processes at unit time intervals, their sample CCF $\hat{\rho}(k)$ gives an estimate of the strength of a possible linear association between the processes behind the observations at each possible lag number k. It is defined as



(a) Regular correlation estimation



(b) Slotted estimator



(c) Weighted estimator

Figure 2.5: Principles of correlation function estimation: (a) shows the classical estimator, where the correlation $\hat{\rho}_{xy}(k)$ is given by a mean over products of zero-mean observations a lag k apart. (b) For irregularly sampled time series, the slotted estimator computes $\hat{\rho}_{xy}(k)$ as the mean over all products in bins whose centers are a lag k apart. (c) Non-rectangular correlation uses the weighted mean over all available products with the weight maxima a lag k apart.

$$\hat{\rho}_{xy}(k) = \hat{\rho}_{xy}(k\Delta\tau) = \hat{\gamma}_{xy}(k)/\hat{\sigma}_x\hat{\sigma}_y \tag{2.9}$$

$$= \frac{1}{\hat{\sigma}_x\hat{\sigma}_y(N-k)} \sum_{t=1}^{N-k} (x_t - \bar{x})(y_{t+k} - \bar{y}) \ . \tag{2.10}$$

Here, $\hat{\gamma}_{xy}(k)$ is the sample cross-covariance at lag k, N is the number of observations, $\hat{\sigma}_x$, $\hat{\sigma}_y$ the sample standard deviations of the processes and $\bar{x}$, $\bar{y}$ are the estimated mean values of the time series [29]. The spacing of the CCF lags, $\Delta\tau$, equals – in this standard definition – that of the time series x_t and y_t, $\Delta\tau = t_i^{x,y} - t_{i+1}^{x,y}$.

The discrete Fourier transform of the sample CCF is the sample cross spectral density function or cross spectrum and vice versa. The power spectrum can thus be estimated in two ways, either by computing the discrete Fourier transforms of the input time series and multiplying them after complex-conjugating one of them, or by estimating the CCF and Fourier transforming it (cf. [29] for more details). All estimators are denoted in the definitions in their respective sections by $\hat{\rho}$, for the sake of simplicity.

The resampling approach for irregular time series

In the case of irregularly sampled time series, the classical definition, as illustrated in Fig. 2.5a, can not be readily applied. An irregularly spaced time series is a pair (t^x, x) of tuples of common length N^x, where $t_1^x < t_2^x < \cdots < t_{N^x}^x$ are the time points and x_i is the value at time t_i^x. For simplicity I have transformed the time variable to get a normalized mean increment of 1 by dividing by the mean sampling period: $t_i^x = t_i^{orig}/\Delta t^x$ and I will use this notation in the following. The differences between observation times $\Delta t_i^x = t_i^x - t_{i-1}^x$ are then not any more constant and the mean of their distribution is the mean sampling time Δt^x. When irregularly sampled time series (t_x, x), (t_y, y) of second-order stationary processes with zero mean are considered, these have to be resampled onto a common regular time grid $(t^{x,y})$ with constant time increments $t^{x,y}(n) - t^{x,y}(n-1) = \Delta t_x$ for all $n = 1, 2, \ldots N^{x,y}$. The grid spacing used in the following is the larger of the mean sampling intervals of the time series.

For brevity this analysis is restricted to the linear interpolation technique, as the effects of other standard routines are not much different in their variance reduction towards the high-frequency end of the spectrum [140]. A resampling method which does not result in a reduction in variance is the *nearest neighbor technique*, where the function is approximated at the desired grid points by the value of the observation closest in time. This leads to a shifting bias [22] which, in the presence of large gaps in the data, can be rather large. I therefore do not employ this scheme. After resampling, the standard FFT-based routines can be employed.

Lomb-Scargle approach

The Lomb-Scargle approach to the spectral estimation of irregularly sampled data can be understood as a least squares fitting of sinusoids to data [134]. The Lomb-Scargle Fourier transform (LSFT)

$$LSFT_x(\omega) = F_0(\omega) \sum_{i=1}^{N^x} (Ax_i \cos\omega\hat{t}_i^x + iBx_i \sin\omega\hat{t}_i^x), \tag{2.11}$$

uses the explicit observation times $\hat{t}_i^x = t_i^x - \tau^x(\omega)$ shifted by the constant phase shift

$$\tau^x(\omega) = \frac{1}{2\omega} \tan^{-1}\left(\sum_i \sin 2\omega t_i^x \Big/ \sum_i \cos 2\omega t_i^x \right), \tag{2.12}$$

to ensure time invariance of the $LSFT$ [140]. The coefficient F_0

$$F_0(\omega) = \frac{1}{\sqrt{2}} \exp(-i\omega t_1^x - \tau^x(\omega)) \tag{2.13}$$

allows for a time shift in the alignment of the two time series in bivariate spectral analysis. The amplitudes A and B are defined as

$$A(\omega) = \left(\sum_i \cos^2 \omega \hat{t}_i^x \right)^{-1/2}, \quad B(\omega) = \left(\sum_i \sin^2 \omega \hat{t}_i^x \right)^{-1/2}. \tag{2.14}$$

In the univariate case, the well-known Lomb-Scargle periodogram is then given by

$$\hat{P}_x(\omega) = LSFT_x(\omega) LSFT_x^*(\omega) \tag{2.15}$$

The (bivariate) cross spectrum can be estimated as

$$\hat{P}_{xy}(\omega) = LSFT_x(\omega) LSFT_y^*(\omega) \tag{2.16}$$

which can be inverted, using the Fourier transform [136], to get the cross-correlation coefficient estimate

$$\hat{\rho}_{xy}(k) = \mathfrak{F}^{-1}[\hat{P}_{xy}(\omega)]. \tag{2.17}$$

The squared absolute value of the $LSFT$ gives the widely known and used LS periodogram [140]. The choice of the frequencies ω is described in [136] and the recommended values for the fundamental frequency $\omega_0 = \omega_{min} = \frac{\pi(N^{xy}-1)}{(t_{max}-t_{min})N^{xy}}$ and maximum frequency $\omega_{max} = \frac{2\pi}{\Delta t^{xy}}$ are adopted. In the bivariate case the observation times $t_{\min}$ and $t_{\max}$ are defined as the lower and upper bounds of the overlapping part of both time series x_t and y_t, otherwise, in the univariate case, minimum and maximum observation time are used. $\Delta t^{xy} = max(\Delta t^x, \Delta t^y)$ is the common sampling rate defined for the bivariate case. The number of frequencies $N_f = \text{ofac} \cdot N^{xy}$ determines the spacing of the frequency vector. According to [67] there is no principal limit, the oversampling factor ofac$>$ 1 is regarded as a smoothing factor, although the number of independent frequencies is constant. I use ofac = 2.

A thorough introduction to bivariate Lomb-Scargle spectral estimation was given by [140]. The use of the technique for correlation function estimation, however, has not yet been explored, though it was already proposed in [136].

Correlation slotting

The sample correlation function $\hat{\rho}_{xy}(k)$ at a lag k is calculated by averaging over the lagged products of the standardized observations. For irregular time series the inter-sampling times vary, and without resampling Eq. 2.9 cannot be applied. An alternative is the *slotting* or *Edelson & Krolik* technique [45, 103]. Its key idea is to calculate the cross-products of all available standardized observations and discretize them into bins according to their sampling time differences as can be seen in Fig.2.5b. The technique was developed in fluid mechanics and applied in astrophysics. $\hat{\rho}(k\Delta\tau)$ at the lag $k\Delta\tau$ is then defined as

$$\hat{\rho}(k \cdot \Delta\tau) = \frac{\sum_{i=1}^{N^x} \sum_{j=1}^{N^y} x_i y_j b_k(t_j^y - t_i^x)}{\sum_{i=1}^{N} \sum_{j=1}^{N} b_k(t_j^y - t_i^x)} \tag{2.18}$$

and the *kernel* $b_k(t_j^y - t_i^x)$ selects the products whose time lag is not further than half the bin width from $k\Delta\tau$:

$$b_k(t_i - t_j) = \begin{cases} 1 & \text{for } ||(t_j - t_i)| - k| < \frac{1}{2} \\ 0 & \text{otherwise .} \end{cases} \tag{2.19}$$

Note that the observations have to be standardized to zero mean and unit variance before the analysis. The lag bin width $\Delta\tau$ is set to be equal to Δt^{xy}, and since the observation times are divided by this mean sampling interval, they can be omitted it in the formulae above, for easier readability (cf. Sect. 2.5.2). I do not choose this width arbitrarily but rather in the context of the desired time resolution of the CCF, more on this in Sect. 2.5.2.

There are, however, several disadvantages of this technique, primarily a high variance of the estimator [4, 10, 64] due to which I will not use this method in the following, but rather apply related, non-rectangular kernels. It also does not always provide positive semidefinite covariance matrix estimates, a problem which can be overcome by 'fourier filtering'. This is discussed further in section 2.5.2.

Non-rectangular kernels

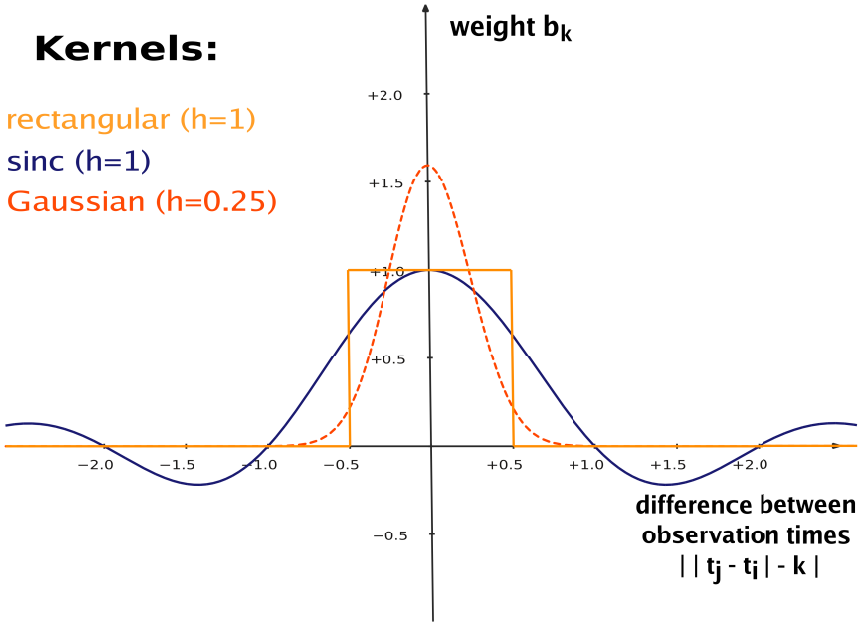


Figure 2.6: Kernel-based estimators effectively 'use' observations whose inter-sampling time difference is close to the lag for which linear correlation is estimated. Slotting (the rectangular kernel) chooses observations within an interval, Gaussian and sinc kernel weigh the products smoothly according to the difference between observation interval and desired lag. Kernels were scaled to the standard choice for width parameter h (cf. Table 2.1, Fig. 2.7).

In analogy to the slotting approach, and taking it further, weighted averaging of the observations can be performed using symmetric, smooth density functions that tend to zero for time differences much larger or smaller than the desired lag k [120]. The similarity is illustrated in Fig. 2.5c. These requirements are for example met by the sinc kernel [154] but also the Gaussian kernel (cf. Table 2.1) as can be seen in Fig. 2.6. Instead of binning the observations into discrete sets, the weights prevent a sudden cutoff in the time domain.

There is no theoretical definition of the effective width of the weight functions. I decide to

Table 2.1: Kernels $b(d)$ used in this work. d denotes the distance between the inter-observation time Δt_{ij}^{xy} and $k\Delta\tau$, k denotes the k-th lag. The standard width parameter h is chosen to result in a main lobe width of Δt^{xy}, the mean sampling interval or common sampling period in the bivariate case.

Kernel [reference]	$b(k - \Delta t_{ij}^{xy}) = b(d)$	**Standard choice for** h
Rectangle [45]	$\begin{cases} 1 & \text{if } d \leq h/2\,, \\ 0 & \text{if } otherwise\,. \end{cases}$	$\Delta t^{xy}/2$
Sinc [154]	$\frac{1}{N}\frac{\sin(\pi h d)}{\pi h d}$	Δt^{xy}
Gaussian [7]	$\frac{1}{\sqrt{2\pi}h} e^{-\lvert d\rvert^2/2h^2}$	$\Delta t^{xy}/4$

scale them to a kernel width of the mean sampling rate for two reasons. i) This choice ensures that – for non-rectangular kernels – observations at (near-)regular times are rated higher than those that are further away, but are still included in cases where little information is available. ii) In a trade-off between the loss of resolution and control of estimator variance, the desired resolution of the correlation function also plays a role, as a kernel width choice larger than the lag spacing would result in mixing information for adjacent lags. The width parameters for the kernels and their relation to the mean sampling rate were confirmed as empirical optima in case of irregular time series (cf. Fig. 2.7). Other parameter choices might, however, also be sensible, depending on the nature of the time series and the statistic to be estimated.

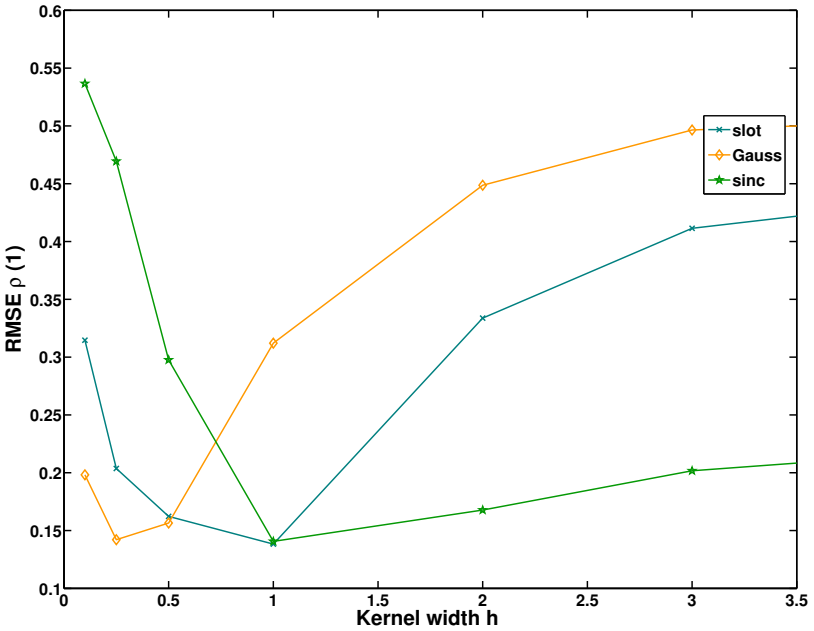


Figure 2.7: Influence of varying kernel width h on the RMSE of $\rho(1)$, using the kernel estimators (cf. Table 2.1, Fig. 2.6). 100 Realizations of sinusoids with random phase in colored noise (30%) were sampled using $\Gamma-$distributed sampling intervals ($sk = 2.85$). (cf. Sect. 2.5.3).

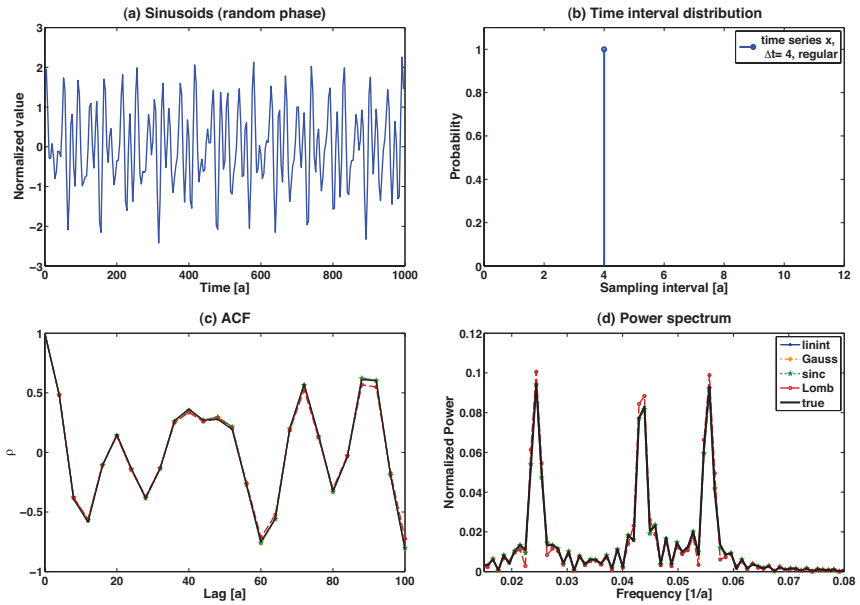


Figure 2.8: Autocorrelation analysis of synthetic signals: For a regularly sampled combination of sinusoids (cf. Eq. 2.21) a sample time series (a), the sampling interval probability density (b), the expected correlation function (c) and the corresponding power spectrum (d) determined from 100 realizations of sinusoid time series with random phase arguments are given. Legends for each row are given in the right panels. All estimators perform equally well.

Positive semidefiniteness of the estimated function

In connection with the slotting-based covariance estimation, the issue with the possible lack of positive semidefiniteness of the correlation estimates has been discussed in [18, 64, 154]. By Bochner's theorem, positive semidefiniteness of the correlation function is necessary and sufficient to ensure non-negativity of the Fourier transform estimate of $\hat{\rho}(t)$. A function $\hat{\rho}(k)$ is positive semidefinite if

$$\iint \hat{\rho}(l-t)w(t)w(l)dt\ dl \geq 0 \tag{2.20}$$

for all integrable functions w, and only if this holds true $\hat{\rho}(k)$ is a possible correlation function. For discrete, short, and regularly sampled time series, using Eq.2.18 and a simple, integrable function for w, this condition is violated for all kernel methods. This problem can, amongst others, be solved by a technique called 'Fourier filtering', which involves Fourier-transforming the correlation function estimate, setting any negative power estimates to zero and applying an inverse FFT afterwards to obtain a positive semidefinite correlation function estimate [4, 120]. Another routine could involve using the absolute value of the power spectrum, instead of setting negative estimates to zero. Also, positive semidefinite matrices have non-negative eigenvalues, which is another means to test this property, and the same modifications as for the power spectra could be applied here. It should be kept in mind, however, that, due to numerical problems, even the 'unbiased' $1/(N-1)$ correlation estimator can result in negative power estimates. When the positive semidefiniteness of the correlation matrix is essential, Fourier filtering should be performed and/or the eigenvalues of the matrix should be checked.

2.5.3 Comparison for synthetic records

To assess the adaptability and suitability of the different estimators, I perform a number of tests on artificially generated discrete signals for which the 'true' ACFs and/or CCFs of the underlying processes are known. For each signal type one first estimates the RMSE in the case of regular sampling. Then time series with Γ-distributed inter-observation times with increasing skewness are created. Since the time vectors are artificial, they do not need to have an actual unit, but I will assume that time is measured in years.

Sinusoids with random phase

Using techniques that are not (yet) fully established, the first concern is to make sure that the results for the standard, regularly sampled case are consistent with those from the standard estimators. Therefore a simple signal, a superposition of three sinusoids, is sampled:

$$x(t) = \sum_{i=1}^{3} \sin(\omega_i t + \Theta_{i,n}) \tag{2.21}$$

with $\omega_i = \frac{2\pi}{T_i}$, $T_i = [18, 21, 41]$ years at a regular rate of 1/4 years. The phase variable $\Theta_{i,n}$ is randomly drawn from a uniform distribution on $(0, 2\pi)$, making this a sample from a stationary stochastic process. The true ACF is then a superposition of cosine functions $\rho_{xx} = 1/2 \sum_{i=1}^{3} \cos(\omega_i)$, irrespective of the relative phases of the signal components. The length of the simulated time series is 1000 years and the function is evaluated for 200 lags. Sample time series, mean ACF and power spectral density (PSD) of the mean ACF are depicted in Fig.2.8. The kernel estimators, the LS periodogram as well as the 'classical' method perform comparably well with a RMSE below 2% (see Fig.2.9, left columns) in the regularly sampled case.

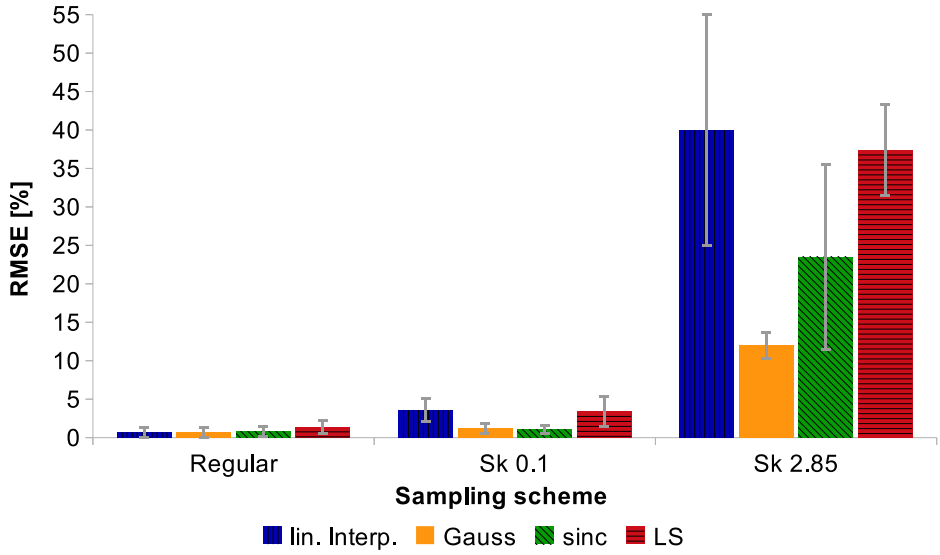


Figure 2.9: Mean RMSE for the ACF estimation (lags 1– 3) using linear interpolation, Gaussian or sinc kernel or the inversion of the Lomb-Scargle periodogram of noise-free sinusoids given for regular, gamma-distributed and mildly irregular (skewness $sk = 0.1$) resp. very irregular ($sk = 2.85$) sampling. Errorbars give the standard deviation of the estimate, calculated using 1000 bootstrap iterations.

I now use irregularly sampled observation times and perform a stepwise increase in sampling distribution skewness (as described in Sect. 2.4.3). For skewness $sk = 0.1$ the RMSEs are only slightly higher (Fig. 2.9, middle columns), but for a skewness $sk = 2.85$ the RMSE is as high as 20% for interpolation and 30% for the LS method. The estimated RMSE for the Gaussian kernel method is rather small compared to that, with an approximate 11%, lower than that of the sinc kernel method (14%). Increasing the skewness in steps of 0.25 from $sk = 0.1$ to $sk = 2.85$ and one can note that the RMSE of the ACF seems to be increasing almost linearly for all the methods. For the LS estimate it jumps in the beginning, from 5% to $\approx$ 20%, and continues to increase at a rate of 9% per unit skewness, with the breakpoint occurring at a skewness of 0.35. The RMSE of the interpolation followed by the FFT-based estimator (denoted 'linint' in the figure legends) increases at a faster overall rate than all the other methods (6.5% per unit skewness). The Gaussian kernel method has the lowest RMSE at high skewness and the lowest increase with respect to the estimate for regular sampling.

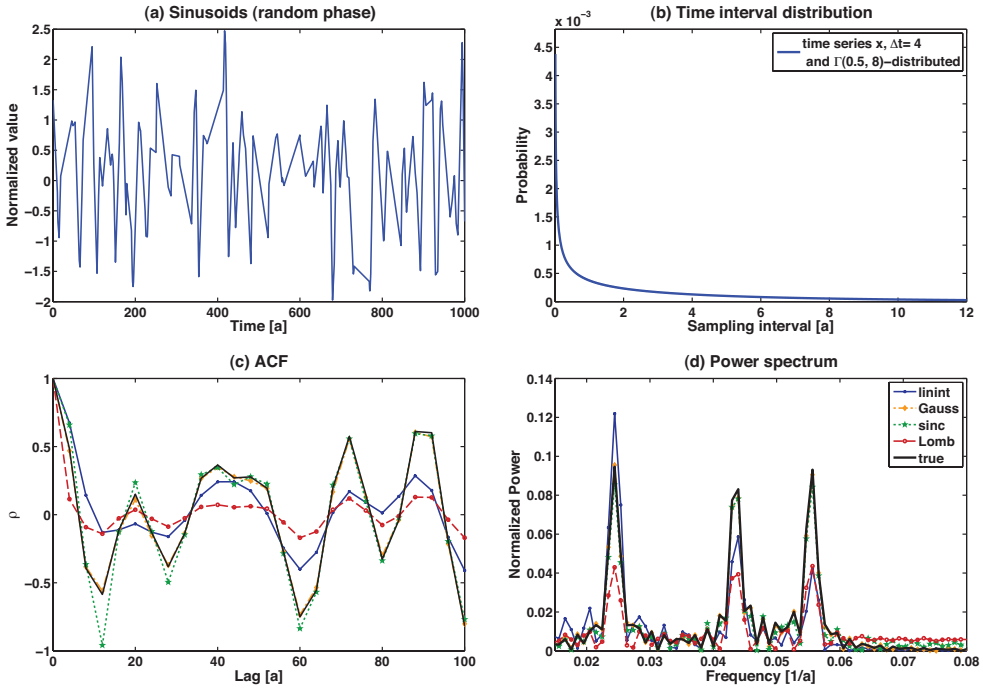


Figure 2.10: Autocorrelation analysis of synthetic signals: For an irregularly sampled combination of sinusoids (cf. Eq. 2.21) sample time series (a), the sampling interval probability density (b), the expected correlation function (c) and the corresponding power spectrum (d) determined from 100 realizations of sinusoid time series with random phase arguments are given. Legends for each row are given in the right panels. High sampling irregularity leads to a variance reduction in the ACF for LS and interpolation.

To investigate the reason for the differences between the methods further, the RMSE of the power spectra obtained from the Fourier-transformed ACFs at the highest input signal frequency $\omega = 2\pi/18$ (c. f. Fig. 2.8d, Fig. 2.10d) is evaluated. I find, that with increasing skewness, the RMSE of this peak increases from around 3% to 10% for interpolation and the LS correlation function estimate, while for sinc and Gaussian kernel it goes from $< 1\%$ to approximately 2%. Estimating the bias of this peak, it is observable that the comparatively high RMSE for interpolation and LS method corresponds to a negative bias increasing linearly from 5% to $> 50\%$ with

respect to the expected peak power at the high-frequency component. In contrast to that, the bias is nearly constant for the kernel methods, the slight increase in RMSE must therefore be due to an increase in variance. This lack of power in the high frequency component of the estimated spectrum is accompanied by a positive bias for the lowest frequency component $\omega = 2\pi/100$ (results not shown).

Autoregressive processes

Understanding the Earth as a complex system with high-dimensional and largely unknown dynamics, dynamical and stochastical approaches to the analysis of the few measurable variables are possible [6]. Since many of the dynamical aspects underlying paleoclimate data are unclear, I will take the stochastic approach. I use AR(1) processes generated at high time resolution and then re-sample the observations onto the desired irregular sampling times. The same simulations as before are performed, first evaluating for regular sampling and then, for gamma-distributed inter-sampling intervals, where subsequently the skewness of the interval distribution is increased. The driving process is given by

$$X(t_i) = \phi X(t_{i-1}) + \xi_i = e^{-\Delta t/\tau} X(t_{i-1}) + \xi_i \tag{2.22}$$

and for bivariate correlation analysis a second process driven by the first at lag ℓ is sampled:

$$Y(t_i) = \alpha X(t_{i-\ell}) + \varepsilon_i \ . \tag{2.23}$$

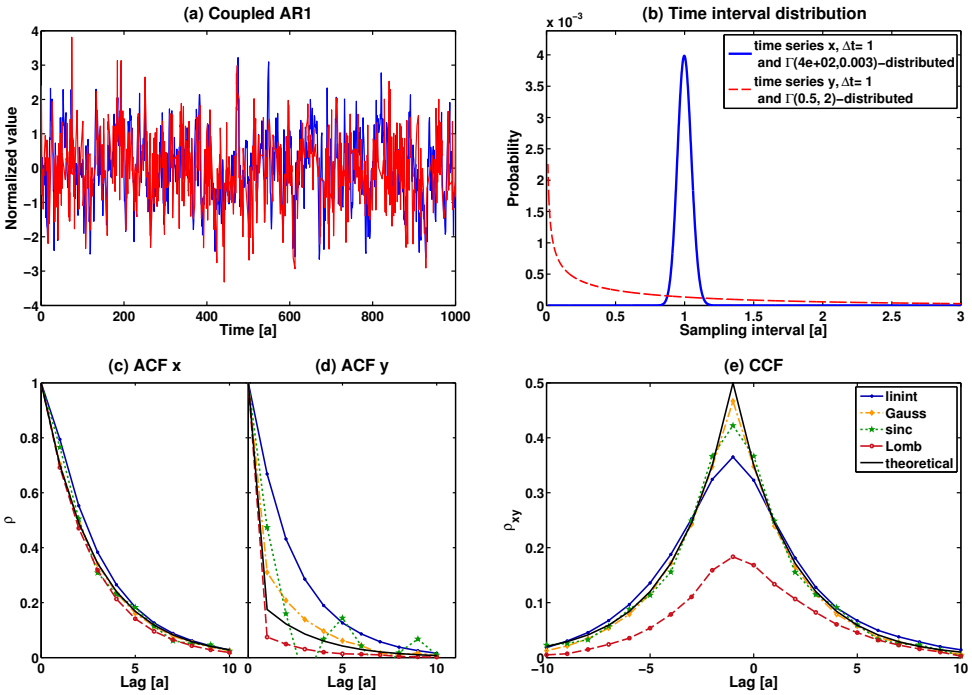


Figure 2.11: Cross correlation analysis for two irregularly sampled signals (cf. Eqs. 2.22, 2.23) from different sampling schemes: Sample time series (a) and sampling time interval histograms (b), the mean ACFs out of 100 realizations (c) and the mean estimated CCF (d). Legends for each row are given in the right panels. A positive bias in interpolation ACF estimates and a negative bias in the interpolation and LS CCF estimates is observable for increased sampling irregularity.

ξ and ε are uncorrelated Gaussian distributed noise processes with a variance σ_ξ^2, σ_ε^2 such that

the overall process variances $\sigma_x^2 = \sigma_\xi^2/(1-\phi^2)$ and $\sigma_y^2 = \sigma_\varepsilon^2 + (\alpha^2\sigma_x^2)$ are equal to unity. I choose the AR(1) coefficient as $\phi = 0.7$, corresponding to a persistence time $\tau = -\Delta t/\ln\phi$, the coupling strength $\alpha = 0.5$, coupling lag $\ell = 1$ and generate the time series (e.g. Fig.2.11a) following the different sampling schemes (e.g. Fig.2.11b). Then ϕ and α are estimated from the time series.

In the estimation of the AR(1) coefficient ϕ the RMSE of the Gaussian kernel and the LS estimator increase only slightly for regular to very irregular sampling from 2% to 4%). The interpolation error increases from 2% to 15% and the error for the sinc-kernel strongly increases for high sampling skewness (from 6% to 12%). I summarize my findings in Fig. 2.12. The RMSE increases subsequently from 2% to almost 15% for interpolation, while Gaussian kernel and LS method remain comparably accurate with an increase from 2% to 4%. The sinc kernel performs not very well, which I think is related to the strong alternating pattern in the estimated ACF bias for time series y_t (cf. Fig.2.11d). The coupling strength α is the true value of the CCF at

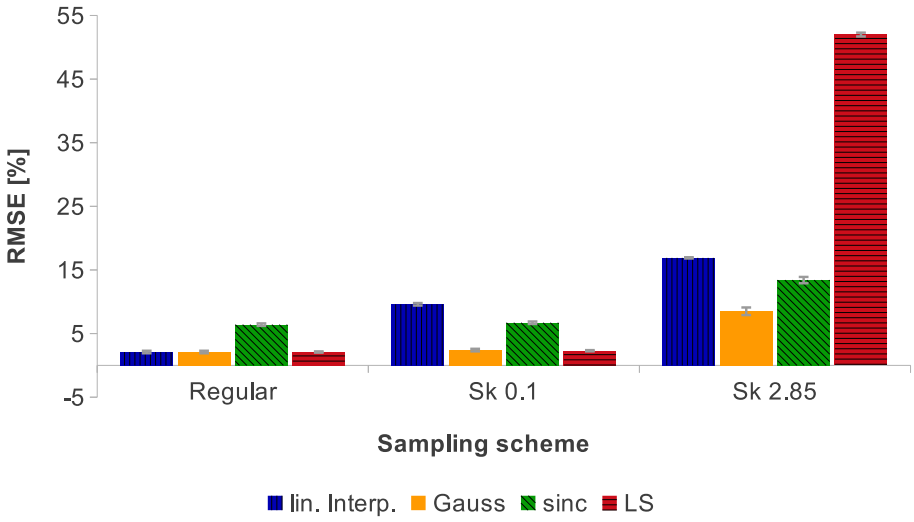


Figure 2.12: Mean RMSE for the ACF estimation (lag 1) using linear interpolation, Gaussian or sinc kernel or the inversion of the LS Periodogram of time series from AR(1) processes (cf. Eqs. 2.22), given for regular, gamma-distributed and mildly irregular (skewness $sk = 0.1$) resp. very irregular ($sk = 2.85$) sampling. Errorbars give the standard deviation of the estimate, calculated using 1000 bootstrap iterations.

the coupling lag ℓ. A typical application in the geoscience context is the estimation of the degree of similarity for time series from different sources with different sampling properties. Analyzing two time series of inter-sampling time distribution skewnesses sk_x =0.1 sk_y =2.85, one finds that the CCF estimation at lag $\ell = -1$ has a negative bias for all techniques. The bias of the LS technique is strongly negative, underestimating the true correlation by more than 65%. Linear interpolation results in a 30% lower estimated coupling strength, the sinc kernel method in 15% and the Gaussian kernel estimate is negatively biased by 8% with respect to the 'true' coupling strength of 0.5 (Fig.2.11e).

Looking at the performance under the increasing sampling time distribution skewness of time series y_t (keeping sk_x constant at 0.1), it is notable that the RMSE of the estimated α increases for all methods, but least for the Gaussian kernel (Fig. 2.13).

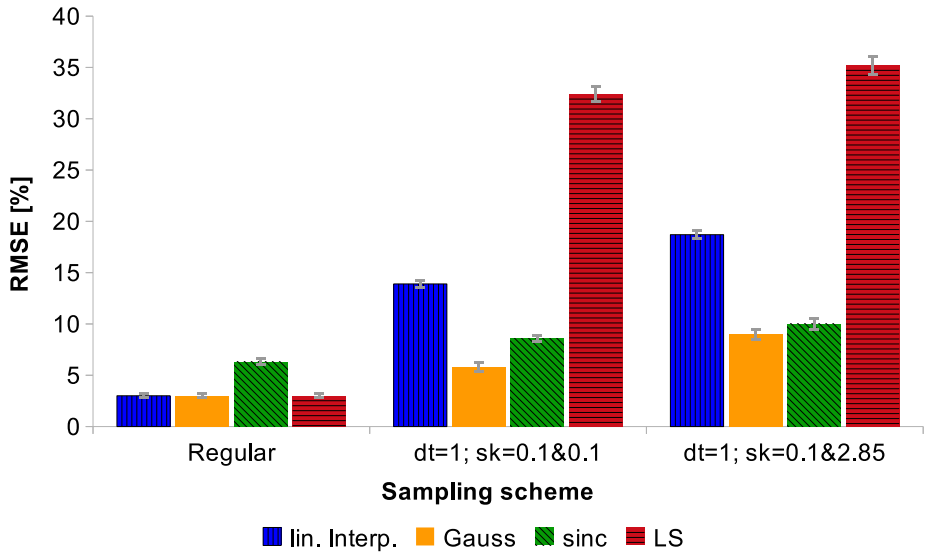


Figure 2.13: RMSE for the CCF estimation (at the lag of coupling) using linear interpolation, Gaussian or sinc kernel or the inversion of the LS Periodogram of time series from coupled AR(1) processes (cf. Eqs. 2.22, 2.23) – given for regular and two gamma-distributed samplings with mildly irregular (skewness sk_x and $sk_y = 0.1$) and very irregular (skewness $sk_x = 0.1$,$sk_y = 2.85$) inter-observation-times. Errorbars give the standard deviation of the estimate, calculated using 1000 bootstrap iterations.

Sinusoids with random phase in colored noise

For irregular time series, the effect of interpolation on the ACF estimation of noise-free sinusoids is that it seems to suppress high-frequency variability. For red-noise signals I find that it, similarly, leads to an overestimation of autocorrelation. To generate more 'realistic' signals, the above-mentioned sinusoidal signals (Eq. 2.21) are synthesized with varying amounts of additive red (AR) noise:

$$x(t) = \frac{1-s}{3} \sum_i \sin(\omega_i t + \Theta_{i,n}) + s\zeta_i \ . \tag{2.24}$$

The sinusoidal components vary with the frequencies $\omega_i = \frac{2\pi}{T_i}$, $T_i = [18, 21, 41]$ years. The time vector t is concatenated into a time line from random variables drawn from a Gamma-distribution with $\mu = 4$ and $sk = 0.1$. The phase variable $\Theta_{i,n}$ is, for each realization n, randomly drawn from a uniform distribution on $(0, 2\pi)$. This makes the time series samples from stationary stochastic processes. ζ_i represents a red noise process (cf. Eq. 2.22) whose variance s I vary in the range $[0, 1]$. The persistence time τ is, for this inter-comparison, fixed at $\tau = 4$ (corresponding to $\phi \approx 0.78$). Since the overall variance of the process is adjusted to equal unity, the signal-to-noise ratio varies in proportion with s.

The 'true' ACF is then given by

$$\rho(k) = (1-s)/3 \sum_i \cos(\omega_i |k|) + s \cdot \exp(-|k|/\tau) \ .$$

Varying s and using irregular time series ($sk = 2$) I find that the mean RMSE of $\hat{\rho}_x(1)$ estimated for the Gaussian kernel method increases slightly from 5% for sinusoidal signals ($s = 0$,

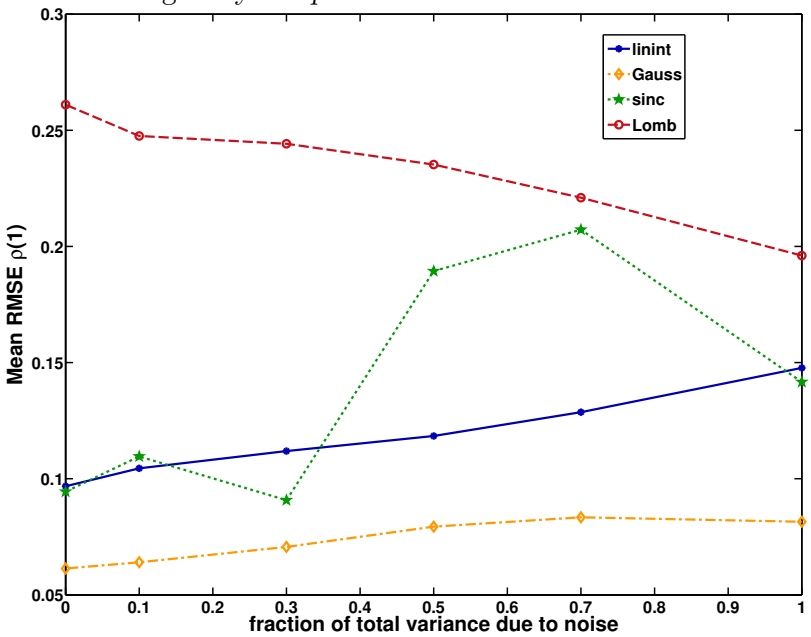


Figure 2.14: Effect of the Signal-to-Noise ratio on the RMSE of the ACF for skewed ($sk = 2$) inter-observation times. The share of the noise variance in the overall process variance increases from left to right (cf. Eq. 2.24).

cf. Fig. 2.14), to 7% for pure red noise ($s = 1$). At the same time, the RMSE for the interpolation-based routine rises from 10% to 15%, that for the LS-technique decreases from 27% to 19%. The sinc kernel performs similar to the interpolation routine for sinusoidal signals with up to 30% of noise, but has a higher RMSE for noise-dominated signals. For irregular time series with low inter-sampling-time distribution skewness ($sk = 0.1$) I find that the RMSE is maximal for medium signal-to-noise ratios, i.e. it is lower for purely deterministic and purely random time series than for the mixture of both (results not shown). For mostly deterministic time series,$s \leqslant 0.5$, the LS technique has then the highest RMSE, while sinc and Gaussian kernel-based methods give more accurate results. For dominant red noise $s \geqslant 0.5$, the LS technique gives good results with low RMSE, where at the same time the performance of the sinc kernel deteriorates. The interpolation-based FFT-routine is not the best choice for irregular time series, irrespective of the signal-to-noise ratios of the processes generating the time series.

The increased RMSE for interpolation observable for the ACF estimates is due to a positive bias for $\rho_x(1)$. The RMSE of the kernel-based methods is lower and the ACF bias is constant and negligible. The high-frequency variability is systematically underestimated when using interpolation. The higher the persistence time τ in the AR(1) component, the lower are the advantages of the Gaussian-kernel based estimator, since the high-frequency variability in the signal is lower.

Summary of the synthetic tests for Correlation estimators

In all tests performed in this section, I find that linear interpolation comes with two systematic effects. Firstly, it has a positive bias for ACF estimation and secondly, it has a negative bias in CCF estimation. Both effects become more severe with increasing sampling time distribution skewness. The LS technique performed well for the ACF estimation of autocorrelated time series but not for sinusoids. The opposite pattern is found for the sinc kernel: its RMSEs are low in the application to sinusoidal data – but high for the ACF of autocorrelated noise processes. The Gaussian kernel estimates are consistent and have the, or close to the, lowest RMSEs in all tests. Therefore I recommend the use of the Gaussian kernel-based estimator instead of – or in addition to – the standard interpolation routine for irregular time series with positive inter-sampling time

distribution skewness, and especially in the presence of observation gaps.

2.6 Mutual information for irregularly sampled time series

Mutual information $I(X,Y) = I_{xy}$ is a measure of the dependency (linear or nonlinear) between two random variables, X and Y. This measure from information theory can be interpreted as the uncertainty reduction in variable X, given that Y was observed. It is symmetric, i.e. relationships of opposite sign but the same association strength, correlation and anti-correlation, give the same MI. By definition, the measure yields a null result if, and only if, the two random variables, in this case time series of observations, are independent [80, 36].

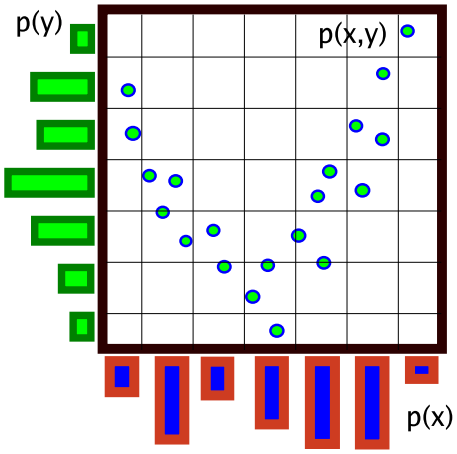


Figure 2.15: Schematic illustration of a simple MI estimation process. The marginal distributions $p(x)$ and $p(y)$ and the joint distribution $p(x,y)$ can be estimated from relative frequencies in 1, resp. 2-dimensional bins.

While more complex estimators exist (e.g. Kraskov et al. [80]), the simplest estimator is

$$\hat{I}_{xy} = \sum_{x,y} p_{x,y} \log \frac{p_{x,y}}{p_x p_y} \ , \tag{2.25}$$

where $p_{x,y}$ is the two-dimensional joint probability density function of the variables X and Y and p_x resp. p_y are the one-dimensional probability distributions of X resp. Y. The unit of measurement of MI depends on the *logarithm* chosen in the estimator: it is measured in *bits*, if the logarithmic base 2 is chosen, and in *nats* for the natural logarithm. Different estimators can be applied to estimate mutual information starting from the marginal and joint distributions of $x(t)$ and $y(t)$. Commonly, these distributions, needed for Eq. 2.25, are estimated from relative frequencies in the binned scatterplot of $x(t)$ vs. $y(t)$, as visualized in Fig. 2.15. In case of irregular sampling, however, the bivariate observation set (X_t, Y_t) at regular observation points t that is required for a scatterplot are not available. In standard interpolation procedures, both (t_x, x) and (t_y, y) would be re-sampled to obtain a bivariate set of observations with regular observation time intervals, (t_r, x_r, y_r). This is undesirable for paleoclimate records a) because every interpolation routine involves an assumption on the dynamics of the underlying process, and this is difficult to justify for climate data and b) it reduces the observable variability in the process [140, 154, 4].

There are two main points where this problem can be addressed: Either by reconstructing bivariate observations while avoiding variance reduction or by a modification of the joint distri-

bution, for example by introducing weights proportional to the sampling time-distance. Following the former solution, the probabilities required for Eq. 2.25 are straightforward to derive from relative frequencies.

I developed and tested several MI estimators with weights on the points in the joint distribution, but found that they were outperformed by an adapted interpolation routine, *local signal reconstruction*:

Algorithmically, this can be described as follows:

1. A local reconstruction of the signal is performed by estimating for each point i in the time series $X = (t^x, x)$ a corresponding observation from $Y = (t^y, y)$, by estimating a local, observation-time weighted mean y_j^{lr} around a time point t_i^x in Y,

$$y_j^{lr} = \sum_{i=1}^{N_y} \mathcal{G}(t_j^x - t_i^y, h) y_i \ , \tag{2.26}$$

   with the Gaussian-kernel based local weight

$$\mathcal{G}(t_j^x - t_i^y, h) = \frac{1}{2\pi h^2} e^{\left(t_j^x - t_i^y\right)^2 / 2h^2} \ . \tag{2.27}$$

   Here, h is the standard deviation of the Gaussian weight function. If there are no observations y_i available in a time window $\pm\tau\Delta t$ around t_i^x this reconstruction is not performed. Repeating this for each time point $j = 1, \ldots, N^x$ in X one obtains a new, bivariate set of observations

$$Y^x = (t_i^x, x_i, y_i^{lr}) \ .$$

2. Afterwards the procedure is repeated by stepping through t_j^y, which yields

$$X^y = (t_j^y, x_j^{lr}, y_j) \ .$$

3. The local reconstruction Y^x and the original observations Y are then concatenated into one vector $Y^r = \{Y \cup Y^x\}$ combining locally reconstructed and original observations. Similarly, a vector $X^r = (X \cup X^y)$ is obtained.

4. Based on this set of bivariate observations (X^r, Y^r) the joint density of X and Y can be estimated using standard binning estimators for MI.

Conceptually, MI is a beautiful method, but it is difficult to estimate in practice, first and foremost because of the large bias effects produced in the inference of the joint and marginal probabilities. Elaborate algorithms have been devised to improve this [80, 119, 132], but no straightforward solution to this has been found yet. I tested several algorithms and finally resorted to the most simple equidistant *binning estimator*, as illustrated in Fig. 2.15. due to its computational efficiency and simplicity. Bias effects are predominantly tied to the temporal sampling and length of the time series due to the occurrence of empty bins. Thus they can, if necessary, be estimated and subtracted using uncorrelated processes with the same observation times as in X and Y. However, for the use as a similarity measure comparable to *XCF* and *ES* in the context of paleoclimate networks I only require that the estimated MI be proportional to the actual association strength. For bivariate normally distributed X and Y *MI* is by definition proportional to the estimated correlation coefficient r_{xy}^2,

$$I_{xy} = -\frac{1}{2} \log(1 - r_{xy}^2) \ , \tag{2.28}$$

and can, by inversion of this equation, be scaled to the positive semidefinite range of the correlation coefficient so that $\hat{I} \in [0, 1]$ [114]. Note that the estimated and transformed value for MI, $\hat{I}$, will only be equal to the estimated value for Pearson correlation $\hat{r}_{xy}$ if the processes are linearly correlated and bivariate normally distributed.

I compared the performance of MI estimation for standard linear interpolation (called *iMI* in the following) and the Gaussian kernel-based reconstruction scheme, denoted *gMI*, at varying sampling irregularities, following the sampling sensitivity analysis described in Sect. 2.5. I generated AR(1) processes at very high time resolution and then re-sampled the observations onto the irregular observation times, as in Eq. 2.22 and Eq. 2.23. Please note that the adapted noise terms in Eq. 2.22 and Eq. 2.23 ensure that the value of the correlation coefficient at the lag of coupling is equal to the coupling parameter α. For the tests in this section I chose $\Phi = 0.5$ and $\alpha = 0.8$ and, at unit average sampling rate, a time series length of 250 units. The expected value for mutual information of these processes at the lag of coupling is given by $MI(X(t), Y(t+l)) = -0.5 \log(1 - r_{xy}^2(l))$, where $r_{xy}(l) = \alpha = 0.8$ is the estimated correlation coefficient, as the used AR(1) processes follow a bivariate normal distribution [114]. Then $I(X(t), Y(t+l))$ can be estimated from the simulated time series and, compared to the expected value, the RMSE of the estimators can be calculated. For the evaluation of the joint and marginal distributions, $n_{bins} = 10$ equidistant bins were employed. In principle, the number of bins should be adapted to the respective length of the time series involved, to reduce bias effects due to empty bins. In the context of paleoclimate networks I am, however, not interested in the value estimated for MI, directly, but only its significance with respect to uncorrelated time series of the same temporal sampling. These are, due to their equivalent sampling, also equivalently biased and therefore the bias effects are negligible in this context.

The results are shown in Fig. 2.16. With increasing sampling irregularity (i.e. larger gaps) the RMSE of the linear interpolation routine increases systematically. This effect is also visible for the Gaussian-kernel based signal reconstruction, but it is much milder. In contrast to *gXCF*, the optimal choice of the bandwidth seems to play a less important role, and kernel widths from $h = 0.25 \ldots 1$ are an equally good choice. I therefore conclude that estimating MI using local Gaussian kernel reconstruction is *more efficient* than using standard interpolation. Nevertheless, if for example the actual value of auto-MI I_{xx} is needed, e.g. for the detection of appropriate time delays in delay embedding [75], more sophisticated tests would need to be carried out to ensure the suitability of the algorithm proposed here, because the removal of bias is not straightforward any longer. In this case an approach similar to that of Albers and Hripcsak [2] could be suitable.

2.7 Event synchronization

The concept of event synchronization (ES) was introduced by Quian Quiroga et al. [123]. The motivation behind the development was to obtain a simple, fast method that quantifies the synchronization between time series where certain *events* can be distinguished. The primary application was focused on neurophysiological signals [123, 81], but it later was also applied for the investigation of rainfall patterns in the Asian Monsoon domain [91, 93].

The main idea behind ES is that two time series are considered to be synchronized if events in time series x occur close in time to events in time series y. Considering the temporal order of the events, e.g. if an event in y occurred *before* one in x, it is also possible to infer which process is *leading*. In the following I will define ES in parallel to the definition given in [123] and [91, 93], and test it in Chapter 3 along with *gXCF*, *iXCF*, *gMI* and *iMI* for irregularly sampled, autocorrelated and possibly time-uncertain time series.

Given two time series (t^x, x) and (t^y, y) that represent observations of autocorrelated stochastic

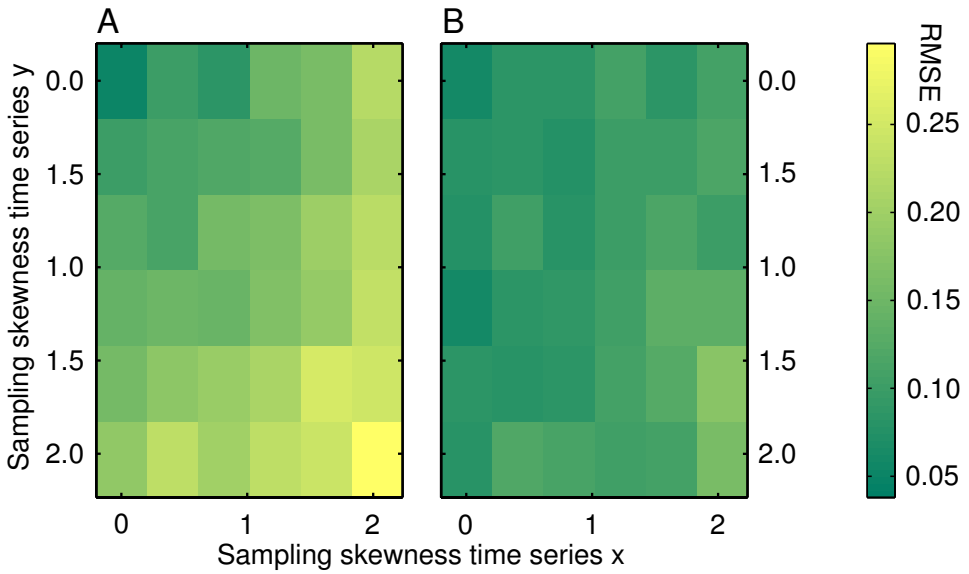


Figure 2.16: Evaluation of the MI estimators for irregularly sampled time series. For each patch on the images 100 coupled AR-processes were generated. Signal construction and sampling irregularity of the time series increases along the x and y axis (analog to [128]). For each pair of time series MI was estimated, (A) based on interpolation to a mean sampling rate and (B) using an adapted Gaussian kernel scheme. Colors indicate the RMSE of the estimated cross-MI at the lag of coupling.

processes, *events* are given by the set of observations that are considered *extreme*, in that their observation value lies above or below the α-th resp. $(1-\alpha)$ percentiles of the distributions of x and y. The actual *value* of the observation at the event points is not relevant for the further analysis. Once the events are defined, only the observation *times* are considered in the event time vectors t_x^* and t_x^*. Next a temporal threshold τ is defined to evaluate the relationship between the events in X and Y with a maximum separation time:

$$\tau = \max\left(\Delta t^x, \min(\Delta t_x^*, \Delta t_y^*)/2\right) \ . \tag{2.29}$$

Here, Δt^x is the mean sampling rate of X, and Δt_x^* and Δt_y^* are the inter-event times in X and Y, respectively.

Subsequently, the co-occurrence of events in X and Y is counted and summed for all events as

$$c(X|Y) = \sum_{l=1}^{N_x} \sum_{m=1}^{N_y} J_{lm}^{xy} \ , \tag{2.30}$$

where N_x and N_y, respectively, give the total numbers of events in X and Y. The counter variable J_{lm}^{xy} is defined as

$$J_{lm}^{xy} = \begin{cases} 1 & \text{if } -\tau < t_l^x - t_m^y < +\tau \\ 1/2 & \text{if } t_l^x - t_m^y = 0 \\ 0 & otherwise\,. \end{cases} \tag{2.31}$$

$c(Y|X)$ is obtained by exchanging X vs. Y in the above expression, and combining both,

$$Q_{xy} = Q_{xy}(X,Y) = \frac{c(X|Y) + c(Y|X)}{\sqrt{N_x, N_y}} \tag{2.32}$$

gives the *strength* of the event synchronization and

$$q_{xy} = \frac{c(X|Y) - c(Y|X)}{\sqrt{N_x, N_y}} \tag{2.33}$$

the *direction* of the association. Unless double-counting of events occurs, these measures are normalized to $-1 \geq Q \leq 1$ resp. $-1 \geq q \leq 1$. $Q = 1$ corresponds to completely synchronous occurrence of events in X and Y, and $q = 1$ implies that all events in Y *precede* those in X.

For the previous studies by Quian Quiroga et al. [123] and Malik et al. [91] local definitions of the temporal threshold τ were used, preventing, in most cases, events from being double-counted, and adapting it to the local inter-event rate. The chosen definition of τ is motivated by the fact that, to be able to compare the results for ES to those obtained from MI and XCF, a similarity function over the *delay* is needed. Thus, the delay τ cannot be arbitrarily large or small, as in Malik et al. [91], Quian Quiroga et al. [123].

I therefore define a ES similarity function based on the proposed measure of event synchronization. It is obtained by shifting the observation times of time series X according to the desired lag:

$$ES(k\Delta t) = Q_{xy}((t_x - k\Delta t, x), (t_y, y)). \tag{2.34}$$

Fig. 2.17 illustrates the transformation of time to event series using the example of the Dandak and Wanxiang cave records. Please note that the definition as a similarity function requires the use of a fixed, or global, event threshold τ, because otherwise the lag time scales are not separable.

2.8 Link strength

In this Chapter I proposed, developed and tested several similarity measures. Each of them comes with different underlying assumptions, estimator bias and variance, and they refer to different properties of the time series: the goodness of a linear fit to the joint distribution (XCF), the sharpness of the joint vs. the marginal distributions (MI) or the relative positions of extreme points, or events, in the time series (ES).

Each of these similarity measures returns estimates whose expectation values are proportional to the actual similarity, despite the sampling irregularity, as I illustrate in Fig. 2.18. To obtain these relationships I used coupled AR(1) processes, see Sect. 2.5.3 for more details. The coupling parameter values α_{true} as in Eqns. 2.22 and 2.23. The expected value of the similarity, α_{est}, and the variance of the estimate are computed as the mean and standard deviations of the estimated $\alpha_{\text{est},i}$ for 1000 realizations.

It is immediately apparent that results obtained from the different estimators are difficult to compare. The MI estimates were converted to the CCF scale and thus are bound to the interval $[0, 1]$. This, together with the substantial and non-negative bias, induces a different proportionality between the actual coupling and the inferred association strength. Inferred ES, on the other hand, increases nonlinearly, but monotonously, with the coupling.

The main use of similarity measures is to assess the association strength between dynamics of processes. This can only be interpreted properly, if the significance of this estimate is known. To unify the results obtained from different similarity estimators I propose to use a *link strength* $p(X,Y)$, which homogenizes and summarizes the results obtained for individual similarity measures.

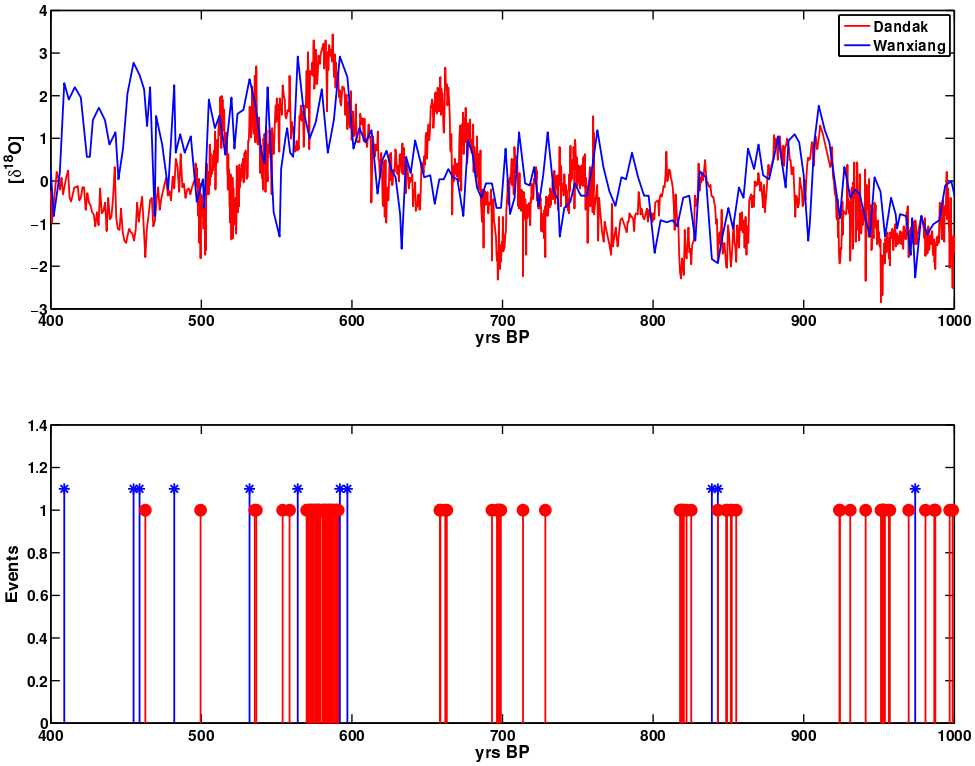


Figure 2.17: Transformation of two given time series to event series for the determination of event synchronization: In the top panel the two time series from Dandak and Wanxiang caves [190, 13] are given. They were normalized but not detrended. They are thresholded using the respective 90% quantiles to obtain the event series, shown below. The observation values of the events are not relevant for the analysis and the bars were plotted at different heights solely for better visibility.

Definition 12 (Link strength) *The link strength $p(X,Y)$ for two observed time series X and Y is defined as the relative frequency of significant estimates out of the N_{sim} employed estimators S_i:*

$$p^q_{sim}(X,Y) = \frac{\sum_{i=1}^{N_{sim}} P_i(X,Y)}{N_{sim}} \ . \tag{2.35}$$

The link strength of the individual estimators, $P^q_i(X,Y)$ is recorded on a binary scale:

$$P^q_i(X,Y) = \begin{cases} 1 & \text{if } S_i \text{ symmetric and } S_i(X,Y) > S^{hi}_i(X,Y) \\ 1 & \text{if } S_i \text{ asymmetric and } \Big(S_i(X,Y) > S^{hi}_i(X,Y)\Big) \,|\, \Big(S_i(X,Y) < S^{lo}_i(X,Y)\Big) \\ 0 & \text{otherwise} \ , \end{cases} \tag{2.36}$$

and here $S^{hi/lo}$ refer to the critical values of a hypothesis test, the null hypothesis being that both X and Y are autocorrelated but mutually uncorrelated, Gaussian distributed stochastic processes. The significance q determines the critical values $S^{hi}_i(X,Y)$ and $S^{lo}_i(X,Y)$ which are obtained from the $q_{hi} = 1 - 0.5q$ and $q_{lo} = 0.5q$ quantiles of surrogate similarity estimates $S_i(X^*,Y^*)$.

In the context of this thesis, independent AR(1) surrogate time series X^* and Y^* are generated on the same time axes as X and Y according to Eq.2.22. The individual AR(1) persistence time is obtained using an efficient least-squares fitting algorithm [128, 108]. I will consider five similarity estimators, *gXCF*, *iXCF*, *gMI*, *iMI* and *ES* in the context of this thesis, but this could be

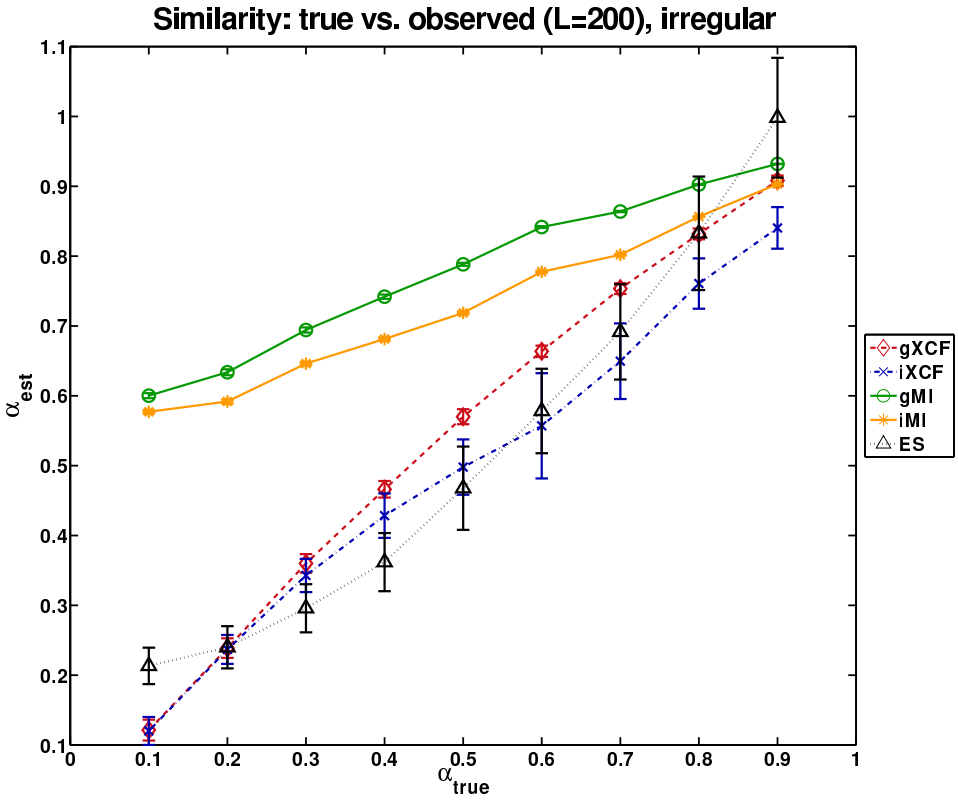


Figure 2.18: Comparison of estimated similarity α_{est} with the known, true similarity α_{true} for irregular time series of length 200, with sampling time distribution skewness $sk = 1.25$. 1000 realizations of coupled AR(1) processes were generated according to Eqns. 2.22 and 2.23. Table 2.2 details the parameter choices for the estimators.

expanded for other concepts, for example based on (cross-)recurrence plots [131, 100, 99, 84], recurrence networks [48], or distance measures [86].

The notion of a link strength, instead of similarity scores, makes it easy to extend the analysis to a whole ensemble of time series, as relevant for paleoclimate time series.

2.9 Summary

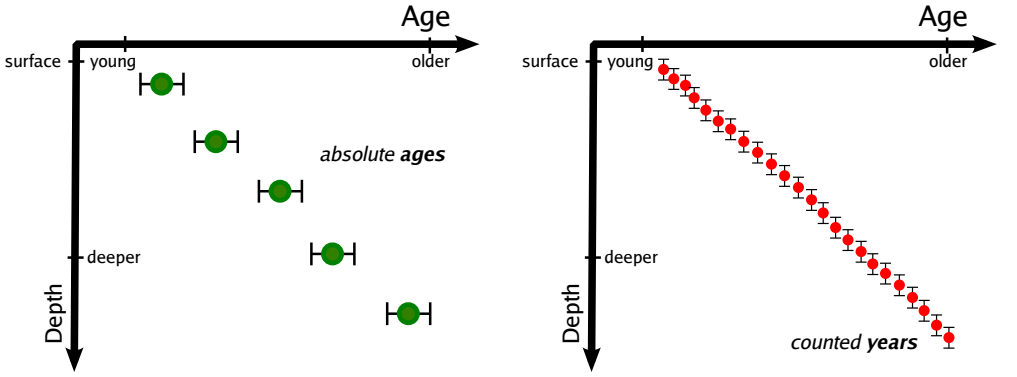
In this chapter I introduced paleoclimate time series and the specific challenges that arise due to the uneven spacing of their observation times. I developed linear (*gXCF*) and nonlinear (*gMI* and *ES*) similarity estimators that do not require regular sampling in time and found that interpolation to regular spacing of the observation times leads to an underestimation of high-frequency variability – and an overestimation of variability on longer time scales. By contrast, the adapted estimators are more efficient in the presence of sampling time irregularity. The significance of results from different estimators and under varying time series length and sampling can be unified using the concept of a link strength. It combines similarity estimators and significance tests and is given by the relative frequency of significant results. Table 2.2 gives a comprehensive overview over the similarity estimators, parameter choices and further references. Chapter 3 tests the proposed similarity estimators and evaluates their robustness in the presence of sampling time irregularity and uncertainty. The link strength concept becomes crucial in the paleoclimate network analysis using ensembles of time series for model (Chapter 3 and 4) and paleoclimate data (Chapter 5).

Table 2.2: Properties, parameters and references of the similarity estimator algorithms for irregularly sampled time series developed and tested in this work.

Estimator (Abbr.)	**Quantif. property**	**Parameter choice**	**References**
1 (*gXCF*)	Gaussian-kernel-based XCF (goodness of linear fit to scatterplot)	$h = 0.25$	this work, [128, 4]
2 (*iXCF*)	interpolation + Pearson correlation (goodness of linear fit to scatterplot)	$\Delta t = \max(\Delta t^x, \Delta t^y)$	e.g. [128, 29]
3 (*gMI*)	Gaussian-kernel-based MI (rel. non-randomness in joint vs. marginal distribution)	$h = 0.5$, $\tau = 3$	this work, [129]
4 (*iMI*)	interpolation + MI (rel. non-randomness in joint vs. marginal distribution)	$\Delta t = \max(\Delta t^x, \Delta t^y)$, $n_{bins} = 10$	[129, 36]
5 (*ESF*)	Relative timing of extreme events	$q = 0.8$	this work, Quian Quiroga et al. [123], Malik et al. [91]

3 Similarity assessment from time series with observation time uncertainty

Paleoclimate reconstructions rely on the measurement of paleoclimate proxies on one hand, and the reconstruction of the time at which the proxy was incorporated into the archive on the other hand. At the first stage, the proxy samples are obtained and recorded against the depth scale, the depth within the archive, relative to the most recent part at the top. *Dating information* is obtained from radiometric (absolute) dating and layer counting (incremental dating), and is depicted in age-depth plots (cf. Fig. 3.1) and summarized in *dating tables*. Absolute correctness of the reconstruction of the *observation time* is impossible, however: Many growth events may be integrated into one proxy measurement, because the final time series resolution depends mainly on the accumulation rate, or growth rate, of the archive. The resulting puzzle for the paleoclimatologist is nontrivial: Given a dating table (relating time of deposition to deposition depth) and proxy measurements (relative to deposition depth), both have to be matched to obtain a time series of proxy observation vs. deposition time useful for climate studies.



(a) Absolute dating

(b) Incremental dating, age-depth plot with depth-uncertainties

Figure 3.1: Illustration of common dating information for paleoclimate archives: Absolute ages (e.g. from radiometric dating) with age-uncertainties (a) and incremental dating (e.g. from layer counting) with depth-uncertainties (b). Often this counting error increases with depth [161], sometimes it is assumed to be constant [170, 182, 158]

To relate the measurements to the age of deposition, a procedure called *age modeling* is employed. Usually, this is done by modeling the accumulation history, or *age-depth relationship* of the archive, at the end of which a *most probable* age-depth relationship is inverted to obtain the most probable deposition times corresponding to the proxy measurements [160, 21, 138]. The uncertainty of the time axis in the resulting time series is usually illustrated by plotting the age estimates with their error bars underneath the proxy measurements, as shown in Fig. 3.2. The drawback of such an approach is that the actual uncertainty in the proxy observations, which is due to uncertainty in the control variable – time – is not directly visible and both are often ignored in numerical analysis of the records [21]. Still, the question remains: "All age models are

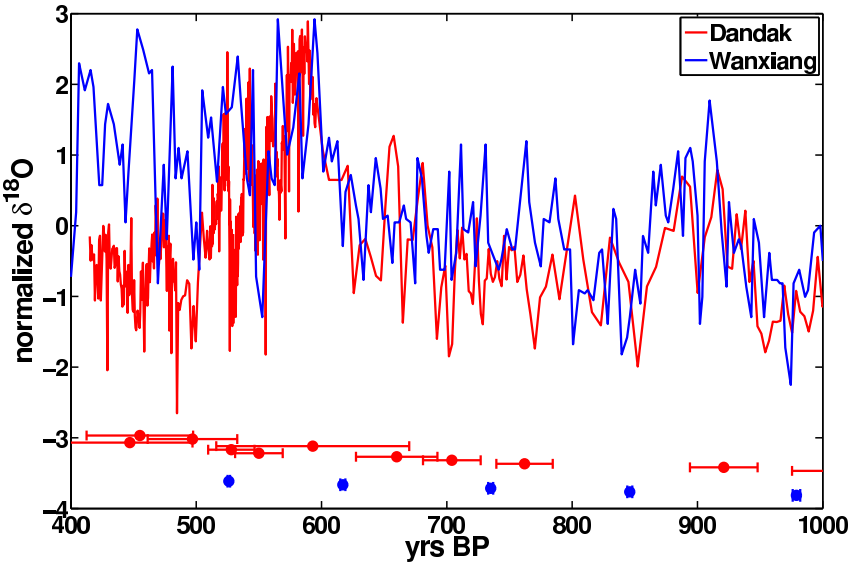


Figure 3.2: To illustrate: Dandak [13, 145] and Wanxiang [190] cave proxy time series as obtained from "classical" age modeling. The respective age uncertainties at the dating depths are indicated by the error bars underneath the time series.

wrong – but how badly? " [160], therefore the impact of age uncertainty on later analyses needs to be assessed.

When comparing two time series, the *relative uncertainty* (c.f. def. 8) between the time series controls the error in the analysis [63], but the *impact* of this uncertainty on the estimated similarity cannot be quantified in visual inspection as illustrated in Fig. 3.2. Standard estimators for statistical similarity require *certain* and *regularly spaced* observation times and are biased in presence of irregular sampling [140, 128]. If the proxy time series is interpolated to regular observation times, and if the uncertainty in the observation times can not incorporated in the analysis, information is *lost* in the process, and a higher degree of confidence in the data is assumed than is warranted.

Therefore this chapter addresses the basic questions:

I) How significant is the impact of age uncertainty on similarity estimation?

II) Are the different similarity estimators for XCF, MI and ES affected in the same way?

III) Given time series of short length and age uncertainty, what is the relative proportion of the different sources of uncertainty (estimator variance, irregular sampling, age uncertainty)?

In the remainder of this chapter I will develop a numerical approach to assessing the uncertainty in similarity estimation and test its effects on short and autocorrelated synthetic time series, similar to those found in paleoclimate applications.

3.1 Approaches to similarity assessment of time-uncertain time series

Uncertainty in input data for some statistical procedure is a common problem in data analysis, as is the desire to have an estimate of the uncertainty in the output data that is due to the uncertainty in the input.

In paleoclimate time series analysis age uncertainty is a key obstacle to be overcome for a comprehensive understanding of Earth system dynamics. To investigate the potential dependence structure of paleoclimate processes X and Y as they are reflected in natural archives, the contribution of age uncertainty to the uncertainty of the similarity $S(X, Y)$ is important.

Thus the aim is to estimate the distribution $\mathbf{p(S(X,Y))}$ of similarity for given datasets X and Y, where

$$X = \left[\mathbb{D}^x = \{D^x, T^x, \sigma_{T^x}\}, Y^d\{d^x, x\}\right] \text{ and} \tag{3.1}$$

$$Y = \left[\mathbb{D}^y = \{D^y, T^y, \sigma_{T^y}\}, X^d = \{d^y, x\}\right] , \tag{3.2}$$

both input datasets consist of a dating table (def. 3) $\mathbb{D}$ with dating depths D, estimated ages T and their uncertainties σ_{T^y} and a set of proxy measurements X^d resp. Y^d (def. 4), visualized as **step 1** in Fig. 3.3. The smoothing resulting from the size of the samples in depth direction, σ_D, is assumed to be negligible here. The input proxy measurements are mapped to observation times in the *age modeling* process (def. 5), **step 2** in Fig. 3.3. In general, algorithms to assess similarity between time series are not capable of processing *probability distributions* or *confidence intervals* instead of singleton values, neither for the observation times nor for the measurement values. In order to get an estimate of the age uncertainty impact on a uni- or bivariate target statistic, these uncertainties have to be included in the analysis. A simple and flexible algorithm feeds many instances of the input data to the statistical estimator to obtain a distribution of statistical estimates that reflects the age uncertainty.

For Pearson correlation, an analytical approach to propagate the uncertainty around the input data into the correlation estimate is feasible. However, Pearson correlation alone is insufficient to characterize similarity between paleoclimate time series in general and in the context of paleoclimate networks. Therefore a Monte-Carlo based approach based on time series ensembles which are obtained via age modeling is used to keep the flexibility regarding similarity estimators.

The task (assessing the uncertainty on the output statistic due to the input uncertainty) can be split into three parts:

Drawing time series from the permitted ensemble of sample ages and corresponding observations (Monte Carlo Simulation, Fig. 3.3 **step 3**), then

analyzing these samples individually, as if they had no age uncertainty (different similarity estimators), **step 4** in Fig. 3.3,

and finally **assessing** the distributions of the output values (i.e. the similarity.), **step 5** in Fig.3.3.

Algorithmically, the approach, illustrated in Fig. 3.3 can be described shortly as

1. In a first step the input datasets X and Y are processed. The monotonicity of the control variables, d and D is checked. If it is not given, in an additional step the dataset is corrected/ modified.

2. A Monte-Carlo simulation for the uncertain age estimates in the dating table is performed: N_{ens} ages are drawn from $T_i^X \pm \sigma_{T_i^X}$ and $T_j^Y \pm \sigma_{T_j^Y}$, respectively, for all $i = 1, \ldots, N_{dtg}^X$ pointwise age estimates corresponding to $j = 1, \ldots, N_{dtg}^Y$ entries in the dating table. This results in dating matrices $\hat{X}$ and $\hat{Y}$ with N_{ens} columns containing the sampled ages. If no distribution of ages is otherwise given the ages are expected to be Gaussian distributed with the given standard deviation.

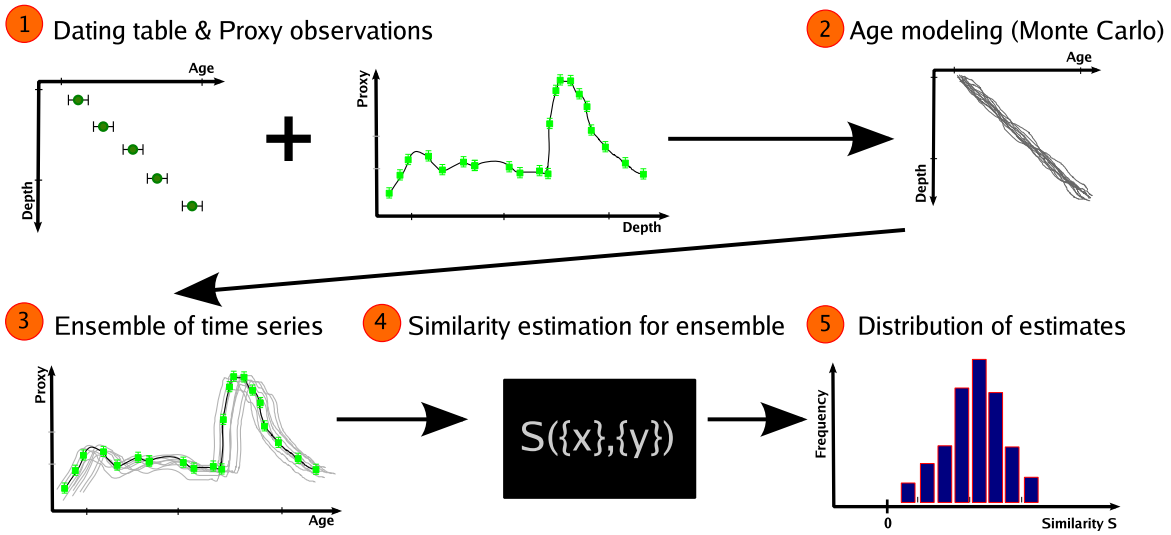


Figure 3.3: Sketch: How much uncertainty is allowed to still enable reliable similarity estimation? Test data with increasing standard deviations of the ages are compared to the estimates for certain observation times.

3. The age estimates in each column and $\hat{X}$ ($\hat{Y}$) are interpolated to the depths of the proxy observations: $T = \texttt{interp}(D, \hat{X}, d)$ which results in a matrix, or an ensemble, of reconstruction observation times $\mathbb{T}$. Together with the depths d, $\mathbb{T}$ forms an ensemble of possible age-depth relationships $\{\mathbb{T}, d\}$ and with the proxy observations x it gives an ensemble of proxy time series $\{\mathbb{T}, x\}$.

4. Each of the members of the ensemble of proxy time series is used as an input to the similarity statistic $S(X, Y)$. This results in a distribution of estimates $p(S(\hat{X}, \hat{Y}))$.

5. Analysis of distribution $S(\hat{X}, \hat{Y})$: Apart from inspection of mean, variance and skewness of this distribution, a hypothesis test can be conducted, comparing $S(\hat{X}, \hat{Y})$ with a distribution obtained from suitable surrogate time series $S(\hat{X}^*, \hat{Y}^*)$.

This approach is general in the sense that it is independent of the specific function $\mathcal{F}([\hat{X}, \hat{Y}])$ that maps the uncertain input to some output estimate. Apart from $\mathcal{F} = S$, $\mathcal{F}$ may represent any bivariate statistic, and with minor modification is also applicable to calculate the influence of sampling uncertainty on univariate statistics, like the autocorrelation coefficients or persistence times [128, 108].

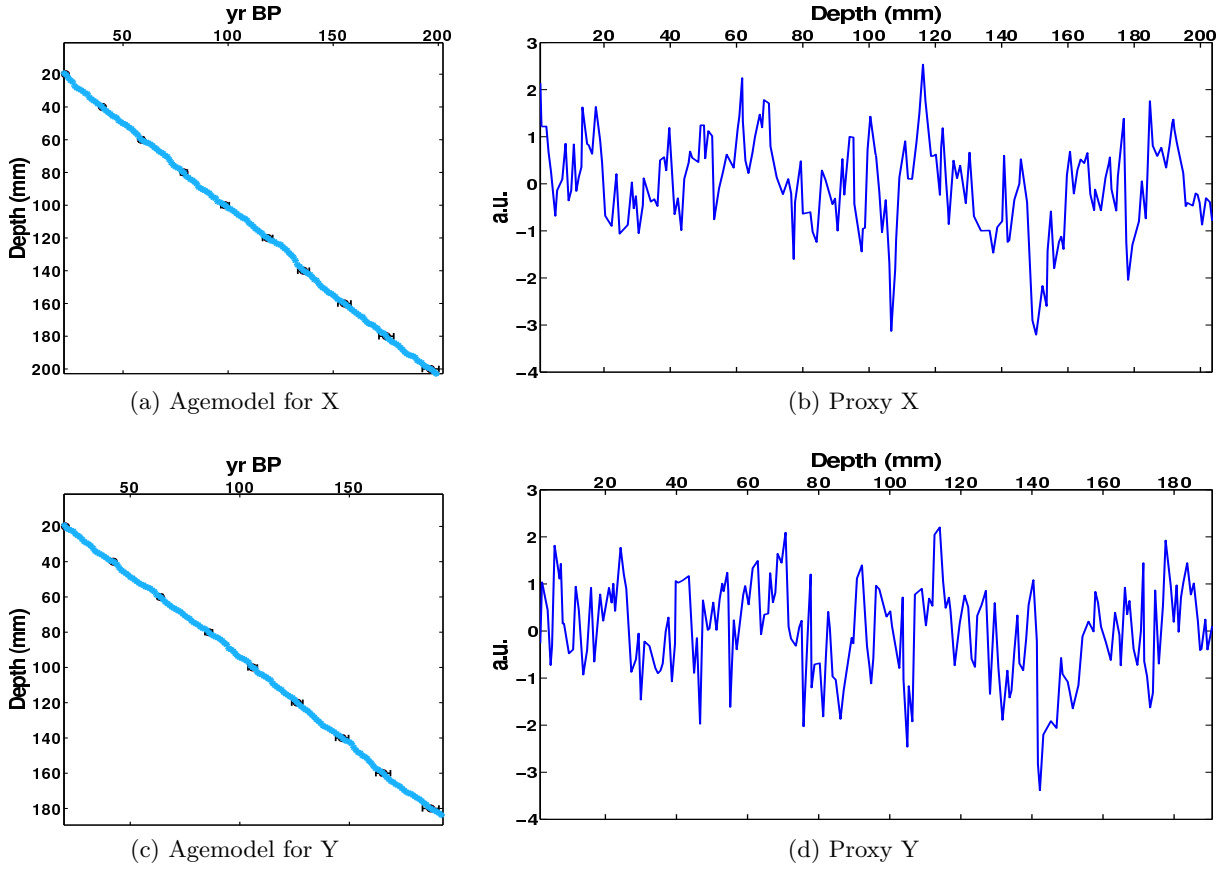
3.2 Sensitivity analysis for synthetic data: How much uncertainty is too much?

This section describes the tests of the MC-based approach to uncertainty estimation on synthetic data. Bivariate similarity assessment is often concerned with estimation of a potential *coupling strength* $S(\ell)$ (hinting towards the same process of origin) and/or the *lag of coupling* ℓ for model-building. For Pearson correlation, the ratio of shared vs. total variance between two processes at a given lag ℓ, $S(\ell)$ is given in the maximum of the cross-correlation function. While the relation to the overall variance of the processes does not necessarily hold by definition for other similarity

measures, they, too, will observe the maximum of their similarity function $\max(\hat{S})$, at the lag of coupling ℓ.

3.2.1 Synthetic data

'True' growth histories for two synthetic stalagmites $SS1$ and $SS2$ and according climate histories are obtained via simulation. These pseudo-archives are then 'dated', and correlated pseudo-proxy for the climate histories are 'sampled'. Then the age modeling procedure is performed and its output is fed into similarity estimation. Finally, I assess, how much of the similarity that was originally present in the climate history is still recognizable significantly, considering the uncertainties. The synthetic data is summarized in Fig. 3.4.



(a) Agemodel for X

(b) Proxy X

(c) Agemodel for Y

(d) Proxy Y

Figure 3.4: Input data for age modeling: Age-depth relationships ((a) and (c)) and proxy records ((b) and (d)) as observations over depth for the synthetic stalagmites $SS1$ and $SS2$. Here, the error in the dating table is set to increase with depth at a rate of 1% (imprecision). The proxy records were generated from a coupled AR(1) process over the *true* ages. The true growth history is given in blue, the point estimates of the age at depth as black dots with corresponding 2σ uncertainty.

The synthetic stalagmite

A synthetic (or: virtual) stalagmite is grown for the sensitivity analysis. The main parameters controlled are

- the growth rate λ in $\frac{\text{mm}}{\text{year}}$,
- the total length of the stalagmite (in mm),
- the type of accumulation (linear growth, or growth modeled via randomly distributed accumulation rates).

As a simple example, a growth rate of $\mu(\lambda(z)) = 1\text{mm/yr}$ is chosen. Linear growth may be a reasonable first order approximation [160], but microscopically, the growth rates of natural archives archive vary. Therefore, $\gamma(\alpha, \mu(\lambda(z))/\alpha)$-distributed accumulation times are drawn for each depth $z_i = \{0, ..., 200\}$mm of the 200mm long stalagmite, with the mean $\mu(\lambda(z))$ determined by the desired growth rate. Please refer to Sect. 2.4.3 for a discussion of the gamma distribution for benchmark tests in paleoclimate time series analysis context. The cumulative sum of the accumulation times then give the 'true' ages of the archive at the depthsz_i: $t_i^{true}(z_i) = \sum_{j=1}^{i} \lambda_i$.

The simulated climate history

As in nature, each synthetic stalagmite $SS1$ and $SS2$ is attached to a climate history. The climate/pseudo-proxy simulation is based on the assumption that $SS1$ lies in an area controlling the climate around $SS2$. Here, climate variability is simulated using coupled AR(1) processes (cf. Sect. 2.5.3 in Chapter 2). The *true* proxy history of climate as recorded in $SS1$ is given by

$$X(t_i^{true}, z_i) = \phi X(t_{i-1}^{true}) + \varepsilon_i, \tag{3.3}$$

and it determines part of the proxy history of $SS2$:

$$Y(t_i^{true}, z_i) = \alpha X(t_{i-\ell}^{true}) + \xi_i. \tag{3.4}$$

Here, ε and ξ are additional Gaussian white noise with unit standard deviation, α is the coupling strength and between $SS1$ and $SS2$ and ϕ the autocorrelation of $SS1$. Since there is no autocorrelative term in Y_t the expected similarity $S(X, Y)$ is equal to the cross correlation of X and Y and the coupling parameter α.

3.2.2 'Dating' of the synthetic stalagmite

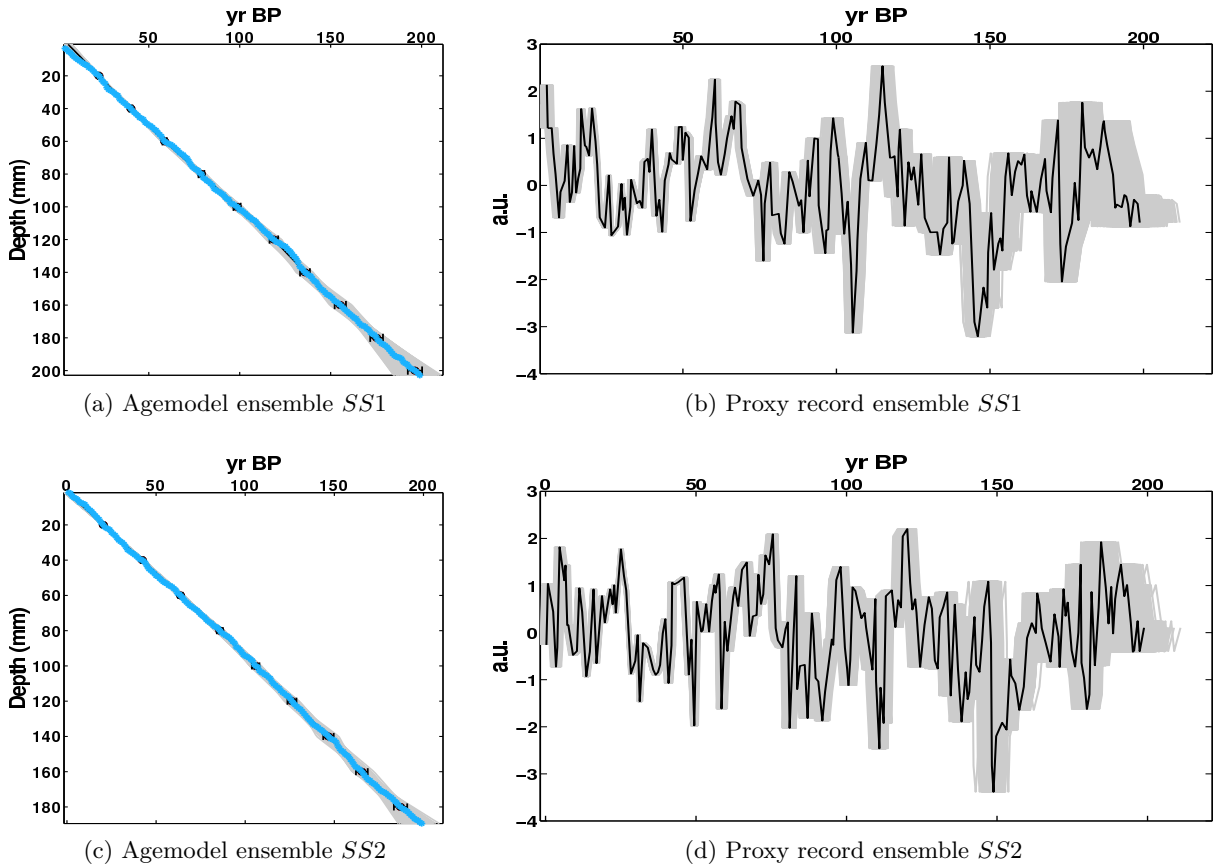
Mimicking the real life situation, the *true* growth history of the synthetic stalagmite, $z(t_{\text{true}})$ is, in the following, inaccessible. The stalagmite is subjected to *dating* along its depth. The dating table contains the for the dating depths D, the estimated age at these depths, T_j, the proxy measurement sample width σ_D and, most importantly, the age uncertainty σ_T.

In real life, the stalagmite would be dated using radiometric dating techniques based on Uranium-Thorium [145, 43, 21] or radiocarbon [182, 179], yielding an estimate of $T(z_j)$ at a few points. The corresponding dating uncertainty, in reality dependent on many factors from initial isotope concentrations, overall age of the core, dating technique to lab and contamination [46], often lies between 0.1 to 0.5% of the age.

For the synthetic stalagmites, dating 'samples' are taken at equidistant depths D_j and the center points of the assumed age distribution are taken directly from the *true* age-depth relationship. The age uncertainty, however, is modeled as increasing proportionally with age, as $p \cdot T_j$. p thus denotes the (im-)precision of the dating and is varied in the following numerical experiments.

3.2.3 Age modeling

Age modeling aims at reconstructing the 'true' depth-age relationship that is inaccessible in real paleoclimate archives. Based on the synthetic stalagmite dating tables $\mathbb{D}^{\curvearrowleft}$ and $\mathbb{D}^{\curvearrowright}$ for $SS1$ and



(a) Agemodel ensemble $SS1$

(b) Proxy record ensemble $SS1$

(c) Agemodel ensemble $SS2$

(d) Proxy record ensemble $SS2$

Figure 3.5: Ensemble of age models (left) and synthetic proxy records (right) for synthetic stalagmite $SS1$ and $SS2$ used in the sensitivity analysis. Here, the error in the dating table is set to increase with depth at a rate of 1% (precision), c.f. Fig. 3.4. Each of the 2000 MC ensemble members is given [cf. [21]].The spread of the proxy values on the right exemplifies that the age uncertainty propagated through age modeling results in blurring of the proxy record towards the bottom of the stalagmites. The *true* growth history, of the archives are given with the light blue markers in the age-depth plots, the median estimate of these ages from COPRA shows in black for both age model and proxy.

$SS2$, the 'observation times' for the proxy observations X^d and Y^d, t^x and t^y are reconstructed by interpolation from the known ages (see Eq. 3.1 and definitions in Sect. 2.3). In Monte-Carlo based numerical frameworks such as StalAge [138] or COPRA [21], an ensemble of age models $\mathbb{T} = \{t_k, z_k\}^{k=1,\dots,N_{ens}}$ is created, which, in their entirety, reflect the age uncertainty of the estimated depth-age relationship. Based on this ensemble of age models, the uncertainty in the similarity estimates can be inferred, as is visible in Fig. 3.5.

In summary, the test plan is thus as follows:

1. Simulate a growth history $z(t)$ of a synthetic stalagmite of length $Z = 200$mm, corresponding to a 'true' age-depth relationship $t_i^{true}(z_i)$, resp. $z_i(t^{true})$. For this, assume gamma-distributed growth and an accumulation rate $\lambda = 1$mm/year.

2. Simulate proxy histories $\{T, x\}^{SS1}$ and $\{T, y\}^{SS2}$ according to the *true* growth history using coupled autoregressive processes (cf. Sect. 2.5.3). Forget the true growth history.

3. Sample the true growth history at the dating depths and infer corresponding uncertainties.

4. Create N_{ens} surrogate dating tables for $SS1$ and $SS2$ with increasing uncertainty of the ages according to the (im)precision p, i.e. an ensemble of dating tables.

5. Assess if the estimates similarity $\mathbb{S}(\hat{X}, \hat{Y})$ is statistically significant for the given uncertainty, and how it is influenced by sampling heterogeneity and time uncertainty.

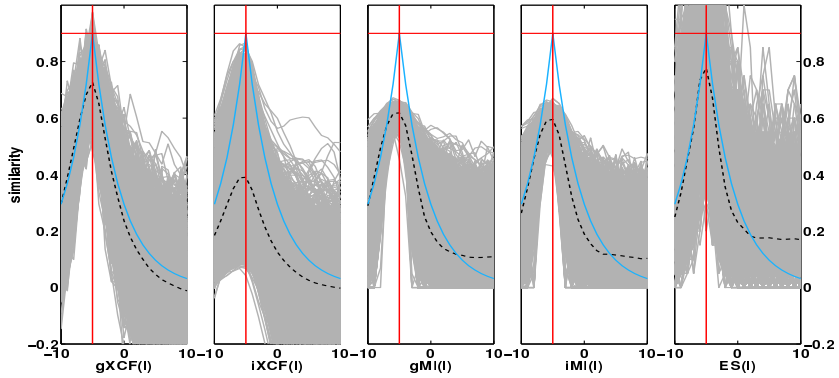
Here, the core of the COPRA algorithm is used for MC simulations. $N_{ens} = 2000$ MC iterations are used to sample the probability space and linear interpolation is employed to infer ages between point estimates of the age at depth.

3.2.4 Results of the sensitivity analysis

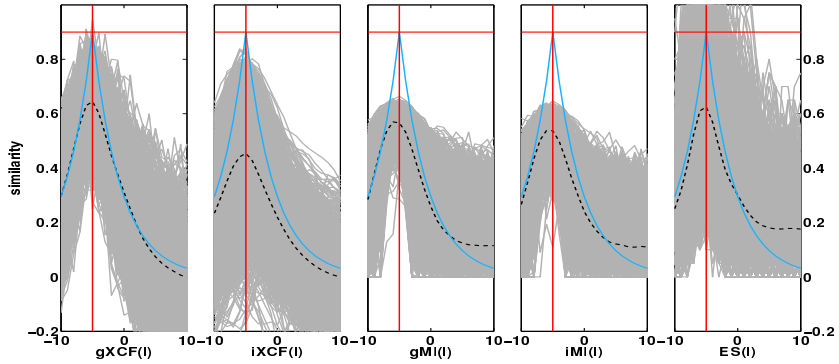
The effect of age uncertainty on different similarity measures, which is in the focus of this chapter, is determined from the distributions of the similarity estimates when compared to their expected value, which is given by the similarity estimate for the corresponding *certain* time series, the 'true' simulated climate history.

Effects of age uncertainty on time series similarity

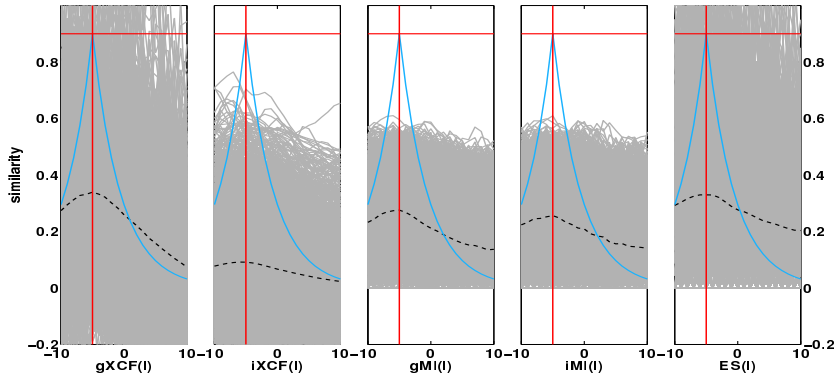
In the analysis of time series, the *time* of an observation implicitly corresponds to an *independent* variable. However, the inter-observation time might *not* be independent of the dependent variable (or the observation value). In the context of time series, statistical similarity refers to similar statistical behavior of observations over time. The independent variable, time, is used to determine which observations in any time series $\{t^x, x\}$ and $\{t^y, y\}$ were made coevally, or at a certain lag time ℓ, and these will contribute to the estimate of similarity $S(\ell)$ at a given lag time $\ell \cdot \Delta t_S$. Age – or time scale – uncertainty blurs the time scale and the assignment of observations $\{t_j^x, x_j; t_k^y, y_k\} \forall j, k : t_j^x - t_k^y \approx \ell \cdot \Delta t_S$ to each other. Given two age uncertain datasets (dating tables & proxy observations) and their uncertain age-depth-relationships mirrored in many realizations of age-uncertain time series, result in an ensemble of similarity functions (Fig. 3.6), and a distribution of similarity estimates $S(\ell)$ at any given lag ℓ. The width of this distribution – and the spread of the similarity estimates – correlates directly with the overall imprecision of the age-depth relationships.



(a) No age uncertainty at dating depths

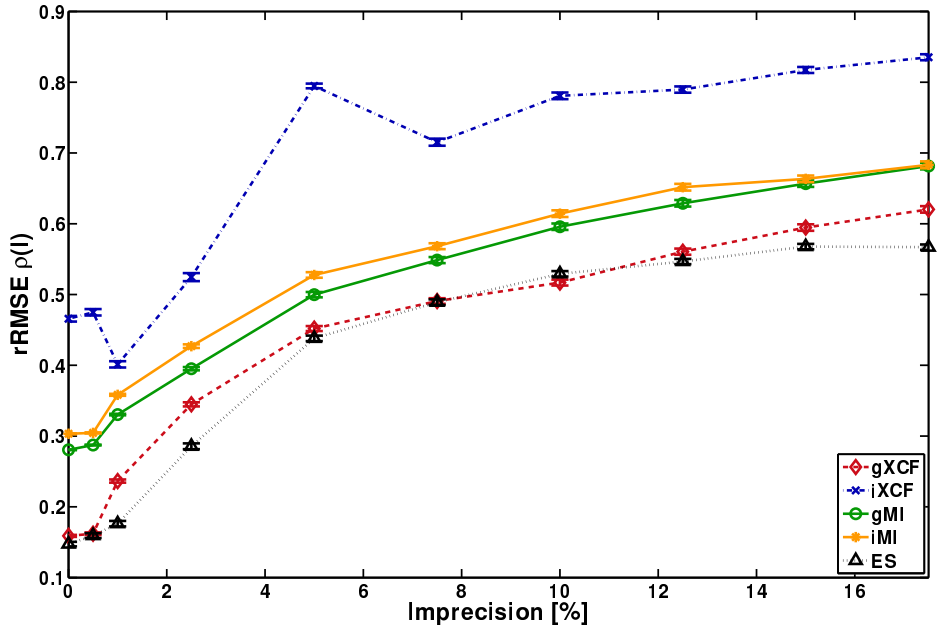


(b) Small age uncertainty (1%))

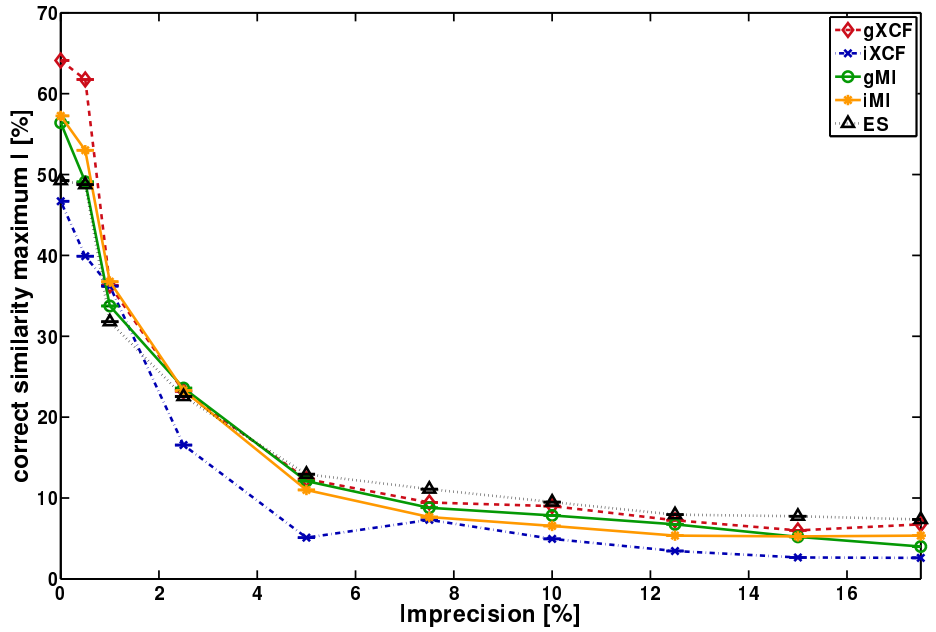


(c) Large age uncertainty (10%))

Figure 3.6: Ensemble of similarity functions estimated from age model ensembles of $SS1$ and $SS2$. The age uncertainty increases from zero in (a) to 1% in (b) to 10% in (c). The age uncertainty is visible in the biased distribution of correlation estimates $S(\ell)$. *True* lag of coupling and coupling strength α are indicated by the red lines and the true similarity function is given in light blue. Dashed lines indicate the median of the similarity function ensemble. Please note that 0% age error does not correspond to a fully certain time axis, as in this case the age is known perfectly at the dating point depths only.



(a) Relative RMSE $\hat{S}(\ell)$



(b) Fraction of correct maximum lag ℓ

Figure 3.7: The relative estimation error $rRMSE$ increases with age uncertainty (a). For coupled AR(1)-processes, the true similarity function has a distinct peak at the lag of coupling. With time scale uncertainty, the percentage of correctly identified maxima (b) drops rapidly at the same time, as the age uncertainty blurs the exact dependence structure. Please note that 0% imprecision does not correspond to a knowledge of the full growth history.

Effectiveness of similarity estimation in presence of errors in the independent variable

Increasing imprecision contained in the time series results in increasing estimation error (i.e. rRMSE) for the similarity at the lag of coupling, $S(\ell)$. The RMSE, especially for MI, is not independent of the 'true' coupling strength $\alpha = 0.9$ (cf. Sect. 3.2.1 and 2.8) and therefore the *relative* RMSE, divided by the coupling strength α, is given in Fig. 3.7a. Note that here the estimator bias unavoidable for MI is estimated using 1000 appropriate AR(1) surrogates and subtracted prior to conversion to the CCF scale, as described in Sect. 2.6.

The time series in the age uncertainty test are generated from AR1 processes (cf. subsection 3.2.1), where process Y is coupled to process X at an intrinsic lag ℓ and with a coupling strength α. For such stochastic processes, the true similarity function is single-peaked, with its peak height determined by α, and its location on the lag-axis by the coupling lag ℓ.

In practical data analysis, the potential lag and strength of (primary) coupling, identified as the maximum of the similarity function is of interest (e.g. for model-building). If no age uncertainty exists (imprecision equal to zero), the maximum of the similarity function correctly identified in 50-60% of the ensemble cases. When time-scale uncertainty exists in the time series, this becomes difficult quickly (Fig. 3.7b). When the percentage has dropped to $\frac{1}{L} \approx 0.05$, where L is the number of lags for which $S(\ell)$ has been estimated, the maxima of the similarity function are perfectly uncorrelated. This limit is approached as an imprecision of more than 10% is reached (cf. the dashed line if Fig. 3.7b).

Error source attribution

The uncertainty around the ages in the dating table is, in MC age-depth modeling, reflected by drawing different 'dates' from distributions around these ages for each MC realization. These realizations will therefore have different partial slopes between any date D_i and D_{i+1}. This corresponds to different estimated growth rates for the individual segments of the synthetic core. At a proxy sampling rate over depth that is constant, this will lead to uneven observation times for the time series which correspond to the MC realizations.

The RMSE of $S(\ell)$ is also dependent on the irregularity of the time series. To separate the sources of uncertainty, $M = 2000$ realizations of coupled AR(1) processes on regular timescales are generated to estimate $RMSE_{\text{reg}}$, the 'baseline' RMSE for each estimator under regular sampling. To estimate the contribution of sampling irregularity to the uncertainty, $RMSE_{\text{irreg}}$, in the similarity estimate for age uncertain time series, the irregular observation time vectors out of the age-depth modeling procedure are used to generate coupled AR(1) processes (see Sect. 2.5.3 for more details). Coupling strength, autocorrelation and time series length are fixed to the previously mentioned values for the three different sampling schemes.

Based on the postulate that the RMSE should increase from regular to irregular to age uncertain time series',

$$RMSE_{\text{reg}} < RMSE_{\text{irreg}} < RMSE_{\text{au}} \ ,$$

the 'baseline' contribution is estimated from regular time series as $RMSE_{\text{reg}}$, that of time scale irregularity as $RMSE_{\text{irreg}} - RMSE_{\text{reg}}$ and the additional RMSE of the age uncertain time series' similarity as $RMSE_{\text{ageunc}} - RMSE_{\text{irreg}}$. The results, averaged over realistic imprecision values, are given in Fig. 3.8. It shows, that the estimators respond differently to age uncertainty. While the estimator bias is low for most estimators, the contribution of increasing irregularity of the time series sampling (due to the uncertain inputs) is non-negligible. The age uncertainty alone accounts for additional, but generally smaller, variability. While a large amount of the uncertainty of the interpolation-based estimators, *iMI* and *iXCF*, is due to sampling irregularity, Event synchronization has a large RMSE for regular time series, that is even higher than that for regular to slightly irregular time series. Therefore the contribution of irregular sampling to

the cumulative uncertainty, as depicted in Fig. 3.8, is negative, thus improving the estimation efficiency!

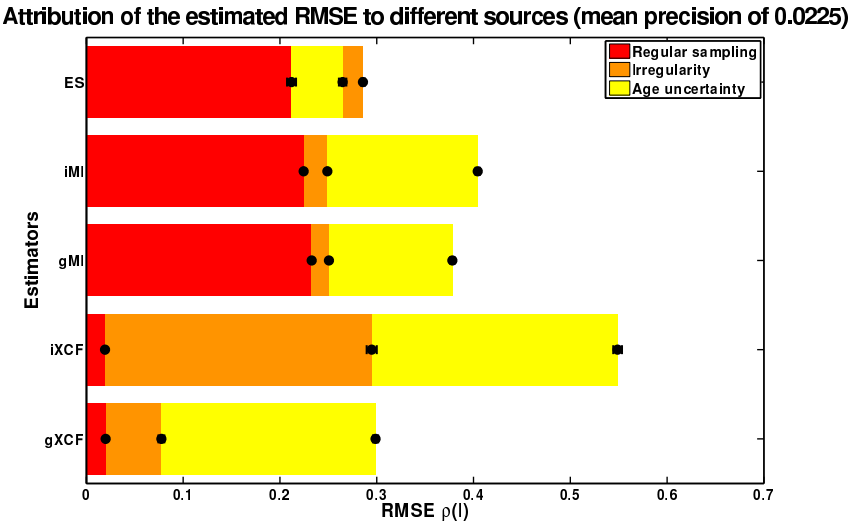


Figure 3.8: Attribution of uncertainty to the sources: General (estimator) error in red, error introduced via irregular sampling (orange) and additional error due to age uncertainty (yellow). The source-dependent RMSE was averaged over the second through to fifth imprecision levels given in Fig. 3.6 as these correspond to the error levels most likely found in real world studies. Errorbars indicate the associated standard deviation. For event synchronization, the RMSE for regular sampling is higher than that for irregular sampling, folding the irregular part of the bar backwards.

3.3 Summary

- Age uncertainty clearly affects all estimators of similarity for time series, and it is an illusion that it would be possible to mitigate the effects of uncertainty on the time axis for any type of analysis depending on observation times. Even if the observation – or accumulation – time of a grown archive is known precisely at some depths, an observation time reconstruction from age modeling requires an assumption on the accumulation behavior which, necessarily, will be wrong to some extent, as nature dictates some extent of stochasticity and irregularity in the growth (cf. Sect. 3.2.4). This is a fact not challenged by the choice of a different interpolation routine, e.g. to a continuous cubic spline, which is often preferred by geoscientists [21, 138]. On the positive side, and although counter-intuitive, incorporating (small) age uncertainty in the analysis might even improve the estimate (cf. Fig. 3.6), when a deterministic (thus: necessarily wrong) assumption on the growth of the archive is made!

- A low imprecision of 0-0.5% or an age uncertainty of approximately 1-2 a.u. over a period of 200 a.u. $\frac{\text{[a.u.]}}{\text{time span in [a.u.]}}$ results in minimal relative estimation error and maximal confidence on the similarity peak position for the time series similarity functions $\hat{S}$. If a similarity analysis for real-world datasets covering a time span of 100 000 years was desired, this would amount to an 'allowed' age error of 500 years at a mean time series resolution of 500

years, which is a lower than what is usually found [159]. Thus, the resolution desired in the analysis is necessarily dependent on age uncertainty – only if that is lower, or comparable, an analysis of such short time series is feasible.

- The similarity estimators tested show, qualitatively, similar performance: The similarity error increases with imprecision and reaches a saturation at about 8-10% relative age error. At this level of imprecision it is also unlikely that the maximum of the similarity function will be found at the correct lag (c.f. Fig. 3.7). At closer inspection, however, the estimators behave quantitatively differently:

  - The **gXCF and iXCF** error split is dominated by the age uncertainty as the largest source of error. Both have small baseline bias for regular sampling. gXCF estimates coupling strength more effectively, however, for both age uncertainty and irregular sampling contributions of iXCF are significantly larger due to interpolation effects.

  - **gMI and iMI** perform badly on the first glance, as their baseline bias for regular sampling RMSE is large. However, one needs to take into account that the RMSE is determined by both variance and bias – and that MI estimation, especially using binning estimators, is always associated with a significant positive bias of the order of 0.6 (re-scaled to correlation units). This bias, however, decreases linearly with the given coupling strength. This bias has been subtracted from the MI estimate prior to scaling it to the correlation scale. The slightly higher baseline bias of gMI is due to the fact that for *regular* sampling the kernel introduces a slight smoothing in the estimation process.

  - **ES**, the measure designed for analysis of event series, performs well and has the lowest total RMSE, followed closely by gXCF. Its baseline RMSE dominates the RMSE split, and the RMSE for irregular sampling is *lower* than that for regular sampling. One reason for this might be that, for irregularly sampled time series of the same mean observation time distance, the number of observations spaced *closely* together is higher, which might increase the chances to find multiple events spaced closely together, resulting in effective *double-counting* of events. The comparably small contribution from age uncertainty indicates, that neither the relative nor the absolute observation time distance between the time series are crucially important to the measure. Thus, it is a more robust similarity measure with respect to age uncertainty than, for example, XCF and MI, which ultimately depend on the notion of simultaneous observations.

4 Spatial dynamics from heterogeneously distributed nodes: Tests with a toy model for Asian Summer Monsoon dynamics.

The development of the **paleoclimate network approach** for the reconstruction of geophysical flows from spatio-temporally discrete and sparse paleoclimate data is at the center of this chapter. Up to now, the estimation of fields of climatic variables (e.g. temperature) or climate extremes (e.g. drought patterns [34, 3]) and links of the climate subsystem to other parameters (e.g. Indian monsoon strength vs. North Atlantic oscillation) have been in the focus of climatologists [176, 144, 178, 69, 62]. Common techniques employed are Spatial Singular Spectrum Analysis (SSA) [176, 55, 78], Empirical Orthogonal Functions (EOFs) [102, 146, 186] and Correlation Maps [24, 174].

Complementary to these approaches, my aim is to extract information on circulation patterns and circulation strength within a climate subsystem from observations at spatially heterogeneously distributed points, as illustrated in Fig. 4.1. In this chapter, I will briefly review the *complex* network approach and the *climate* network approach that emerged from it recently [40, 166, 92, 58, 60]. Based on this I develop the *paleoclimate* network approach, integrating the similarity measures for irregularly sampled time series from Chapter 2 and incorporating proxy record age uncertainty (c.f. Chapter 3). A first test case for the proposed method is the reconstruction of Asian Monsoon dynamics during the last millennium. To evaluate which complex network measures could be suitable to infer the underlying flow dynamics and to what extent spatially heterogeneous sampling affects these measures, I develop and use the semi-empirical simplified model of the Asian Summer Monsoon dynamics **KIMONO**[1].

Key questions:

- Can the paleoclimate network approach be used to reconstruct regional climatic changes?
- Is it possible to extract inter-regional information flow?
- How does spatial heterogeneity affect the reconstruction of spatial dynamics?

4.1 The Paleoclimate network approach (PAN)

4.1.1 Complex networks and the climate network approach

Climate networks are a relatively new tool to explore spatio-temporal variability of climatic parameters and assess dynamical information flow between spatially distant regions [40, 42, 91] and the stability of the climate system and its teleconnections [58, 150, 166, 183]. They are inspired by complex networks theory, which, from sociological through gene regulatory to citation networks consist of two main components: *nodes*, or vertices and *links*, also called edges. The

[1] The name *Kimono* is an acronym that refers to *mo*nsoon and the first two letters of the first names of myself (*Ki*) and *No*ra Molkenthin, who is the mastermind behind the approximation of the Advection-Diffusion equation (c.f. Appendix 1).

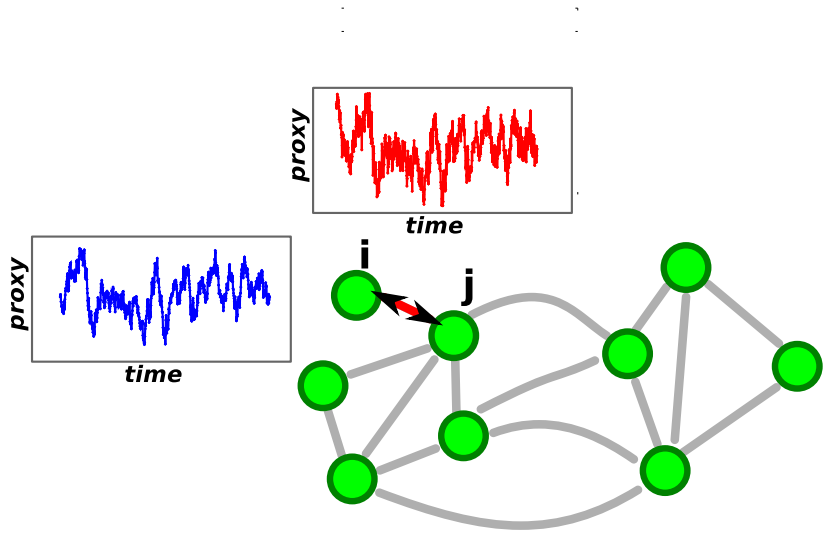


Figure 4.1: Schematic illustration of a paleoclimate network: Nodes in the network represent paleoclimate archives with proxy time series. If the time series similarity estimate between a node pair is significant, these two nodes are considered to be linked. The connectedness of the obtained graph reflects the underlying dependencies created by climate processes.

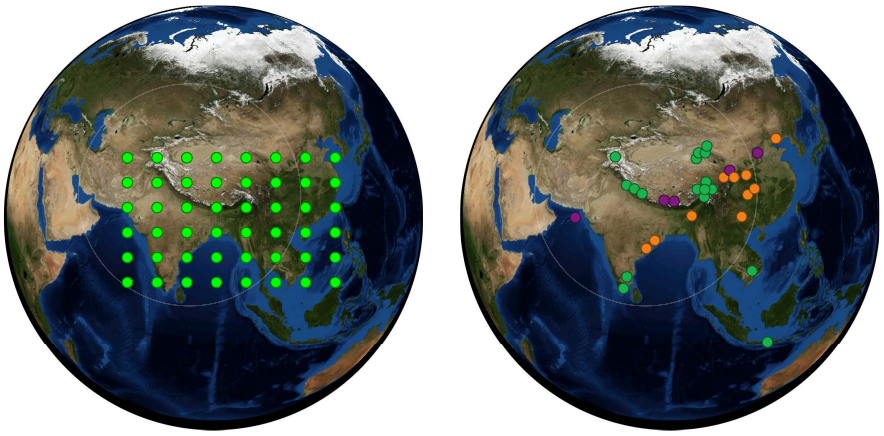
nodes might be representing actors, genes, or authors of scientific papers. The links can be drawn from co-starring in the same movie, sequential expression of genes, or co-authorships.

Climate networks are based on observations of climate dynamics (time series) at certain points on Earth, the nodes (cf. Fig. 4.2a). Computed from these time series, pairwise similarity calculation (linear correlation or nonlinear interrelations, like mutual information (MI) [40] or recurrence-based measures [48]) yield a correlation matrix with entries for each pair of nodes. This matrix is then thresholded using either a fixed value for the correlation or a prescribed *link density*. The resultant *adjacency matrix* $\mathbf{A}$ is a sparse binary matrix with the *(i,j)*th entry being non-zero if (and only if) the time series representing nodes i and j are *significantly* associated. Network statistics can then reflect global and local characteristics of the underlying data: The importance of a node, for example, can be measured by its degree, i.e. how many links the individual node has, or more abstract measures such as betweenness centrality [40, 91].

Previous climate network studies focused on the analysis of gridded datasets, from reanalysis data [40, 42, 58, 150, 166, 183] or recent observations [54, 91, 92], thus they were restricted to the recent, observational period. Palaeoclimate records are, in contrast, spatio-temporally inhomogeneously distributed.

4.1.2 Definition of a paleoclimate network

Briefly speaking, a paleoclimate network is a tool to visualize and analyze dependencies in a given paleoclimate dataset. A schematic illustration of such a paleoclimate network is given in Fig. 4.1. Its nodes are given by paleoclimate proxy archives, its links by significant statistical association between the archives‘ time series. As a method, it asks the question “How dependent are the climate changes in place A on climate changes in another place B – and vice versa?”, rather than “Was the temperature in A strongly linearly correlated to temperature in B at the same point in time?” for EOF analysis. If, say, a proxy for local temperature in China co-varied significantly with reconstructed rainfall in India, this is caused by either a) a common driving



(a) Climate network (b) Paleoclimate Network

Figure 4.2: The climate network approach is based on gridded and dense observations of climatic parameters such as temperature or precipitation (a). Each node in the climate network represents a time series of the same parameter and length. Inspired by this method the paleoclimate network (b) combines paleoclimate proxy records that are heterogeneously and sparsely distributed on the Earth's surface. Nodes may stem from different archives, proxies may reflect different climatic parameters and the temporal resolution of the time series may differ as well.

phenomenon (for example through the North Atlantic Oscillation [181] or solar forcing [1]) or b) local convective phenomena (for example internal ASM dynamics [177]) or c) an artifact in the reconstruction, for example non-climate related common trends in the time series (e.g.ỹouth bias for trees [71].

Consider a paleoclimate network as a graph $G = (V, E)$ on a set of N vertices or *nodes* V, which are connected by a set of edges, or *links* E.

Node

The *nodes* in graph G are embedded in space. Nodes have certain differing *properties* which include their *position* on the Earth surface, the corresponding *type* of paleoclimate archive (e.g.t̃ree, stalagmite, marine sediment) and *proxy* type (e.g.ĩsotope ratios, lithogenic grain size or annual ring width). Conceptually the node location is not relevant. Still, there exist time series whose archive sources are a) distributed over a large area [186], or are b) not Earth-bound, as for insolation [151]. Pragmatically, in the former case, the nodes should be placed in the center of the considered region. The latter should be considered as a node in a different *subnetwork* (see illustration in Fig. 4.3 and the paragraph on subnetworks below.)

Technically, each node is associated with at least one time series of an environmental proxy. If age uncertainty exists and is considered, an ensemble of observation times vectors T_i, $i = 1, \ldots, N_{ens}$ are considered equally likely realizations of the proxy time series of this node. For modeled network results, an ensemble of model realizations for the observation times t can be considered.

Link

In the considered paleoclimate networks, links are *undirected* and *weighted*. A link between node i and node j exists, if the link *strength* is greater than zero, the link weight is given by the link strength.

Fundamental Adjacency Matrix

Fundamentally, for each pair of nodes i and j, time series ensemble member k and similarity measure l a similarity $S_{i,j}^{l,k}$ is calculated. N_{sur} autocorrelated, but mutually uncorrelated time series surrogates are employed to obtain a distribution of surrogate similarity values $S_{i,j}^{*l,k}$. The *Fundamental adjacency matrix* entry $A_{i,j}^{l,k}$ consequently results from thresholding the similarity value $S_{i,j}^{l,k}$ using the chosen critical values $S^*(q_{low})$ and $S^*(q_{hi})$, corresponding to the quantiles $q_{low} = .05$ and $q_{hi} = .95$ of the distribution $S_{i,j}^{*l,k}$:

$$A_{i,j}^{l,k} = 1 \text{ iff } S_{i,j}^{l,k} < S^*(q_{low}) \text{ or } S_{i,j}^{l,k} > S^*(q_{hi}) \tag{4.1}$$

for asymmetric measures that distinguish between positive and negative similarity (here: the correlation-based measures $gXCF$ and $iXCF$), and

$$A_{i,j}^{l,k} = 1 \text{ iff } S_{i,j}^{l,k} > S^*(q_{hi}) \tag{4.2}$$

for symmetric measures that consider only an association strength (e.g. MI and ES).

Link weight matrix

The weight of a link between nodes i and j is given by the ratio of the number of significant associations between them for all N_{ens} ensemble realizations and N_{sim} similarity measures:

$$LW(i,j) = \frac{\sum_{k=1}^{N_{sim}} \sum_{l=1}^{N_{ens}} A_{i,j}^{l,k}}{\sum_{k=1}^{N_{ens}} \sum_{l=1}^{N_{sim}} 1} \tag{4.3}$$

$$= \frac{\sum_{k=1}^{N_{sim}} \sum_{l=1}^{N_{ens}} A_{i,j}^{l,k}}{N_{ens} \cdot N_{sim}} \tag{4.4}$$

Subnetworks

Nodes in the paleoclimate network have different properties (e.g. archive type or geographic origin from a specific region) which may influence its role within the network. To be able to investigate regional dynamics, these nodes can be considered as lying in different *subnetworks*.
A subnetwork H_1 is formed by a subset of nodes $V(H_1)$ and links $E(H_1)$ from network G, where all nodes in $V(H_1)$ fulfill a certain property (e.g. geographic location in the region of 60 – 100° Eastern longitude and 0 – 40° Northern latitude) and all links are between these nodes. Subnetworks are necessarily disjoint, and the unity of the nodes of all subnetworks is the complete set of nodes in the network $\cup_i V(H_i) = V(G)$. If, and how, the nodes are sorted into subnetworks depends on the research question that is asked. For example, splitting the domain as above is motivated by the different ASM subsystems that are thought to influence the regions West and East of this artificial boundary. To investigate the dependency of the ASM on solar irradiance, a proxy record of insolation (e.g. [151]) would then be considered as a node in a separate subnetwork.

Links within the subnetwork, $E(H_1)$ are *internal links*, and links from nodes $V(H_1)$ to nodes in another subnetwork H_2, $V(H_2)$ are *cross-links* $E(H_1, H_2)$. Thus, the overall link set is the unity of internal and cross links amongst the subnetworks, $E(G) = \cup_{i,j} E(H_i, H_j)$. This definition is analogous to that in the *interacting network* approach used in [42] to investigate the relationships between climatic parameters in different atmospheric layers.

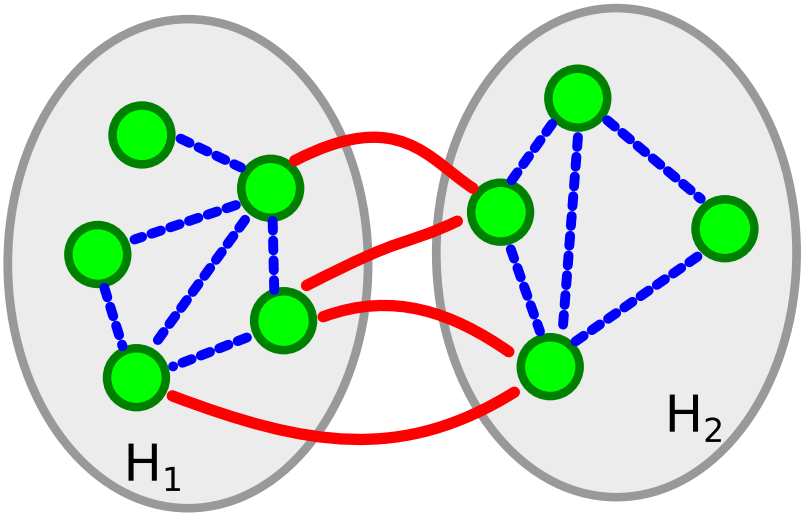


Figure 4.3: To test (paleo) climatic hypotheses, the considered domain is split and nodes in the different subdomains belong to different subnetworks. Statistically, this offers the possibility to differentiate statistically between associations within subnetworks (autolinks, in blue) and those connecting different subdomains (cross-links, in red).

4.1.3 PAN construction

Direct information about past climates beyond the instrumental period relies on paleoclimate proxies. While documentary proxies (e.g. harvest, drought or flood reports for ancient times [24, 29, 74, 117]) are a valuable source of information, only time series of paleoclimate proxy records from natural paleoclimate archives are considered here[2].

Nodes in the paleoclimate network represent a paleoclimate archive, and an associated time series X. Depending upon the type of archive, the node may also be associated with an *ensemble* of time series $\mathbf{X}$, representing for example different realizations of the corresponding depth-age relationship (cf. 3). The spatial distribution of nodes is also time-dependent: While high-resolution archives (such as trees [3, 133, 143], but increasingly also stalagmites [77, 70, 157, 49]) predominantly cover more recent periods at annual time scales, others (e.g. marine or lacustrine sediment sequences, stalagmites or ice cores) might grow at very slow rates, and over the course of millennia [43, 130, 50, 178]. A node will only be incorporated in the network in a time window W if it fulfills the minimal sampling requirements of > 50 observations per window.

Links in the paleoclimate network approach can not be assigned by simple thresholding of a similarity value $S(X, Y)$. Significant similarity is therefore established using surrogate data. The surrogate time series have the same temporal resolution as the original time series, but the observed proxy values are replaced using autocorrelated noise (c.f. Sect. 2.5.3, esp. Equ. 2.22 and [128]). Using n_{sim} different similarity measures S with different characteristics and algorithms

[2]However, the drought/flood reconstruction for China by [186] which I considered in Chapter 5 used historical documents in addition to paleoclimate archive data.

increases the robustness of the link detection, as the proxy-climate relationship might be nonlinear, weak or even erratic. The adapted similarity measures (Gaussian kernel-based correlation and MI estimation: *gXCF* and *gMI* and *ES*, c.f. Chapter 2), form the basis for a network analysis of paleoclimate records, because with these the sampling bias associated with the true dependency structure will remain small. Network construction, as visualized in Fig. 4.1 is conducted according to the following steps:

1. **Building an appropriate database:** In the first step, paleoclimate records in the study region, representing, presumably, one climatic component (e.g. monsoonal rainfall amounts) are identified and checked for comparability: While their time axes does not have to be equal, the average sampling interval should be of the same order of magnitude. Within the time slice of interest, the record should consist of at least N_{minobs} observations, to ensure the power of the similarity tests. For this study $N_{minobs} = 50$ was chosen, which yields a minimal sampling rate of 1 observation per 6 years over the time window.

2. In the second step the suitable **datasets are pre-processed**. The time series are limited to a time window of width W. The nonlinear trend estimated by applying a Gaussian kernel smoother of a bandwidth of $W/2$ is subtracted from each record. The bandwidth is chosen such that potential centennial-scale trends are removed, while high-frequency (annual to decadal) variability is retained. Within each time window the datasets are standardized to zero mean and unit variance. Here $W = 300$ was chosen.

3. In the third step, the degree of **similarity is estimated** for all pairwise combinations of nodes. Within the overlap of the individual pairs, lagged MI and Pearson correlation is calculated in the 'standard' way, involving interpolation to an average time scale, *iXCF* and *iMI*, and using the adapted estimators, *gXCF* and *gMI* as well as the complementary *ES*. As a result n_{sim} matrices with similarity estimates are obtained. For this study $n_{sim} = 5$ similarity estimators were considered: *gXCF*, *iXCF*, *gMI*, *iMI* and *ES*. Please refer to Chapter 2 for more information on the estimators and Tab. 2.2 for parameter choices.

4. Pairwise **significance tests** for each similarity measure S are conducted as described in Chapter 2 and [128]: Surrogate time series are constructed following the null hypothesis that both records are uncoupled irregularly sampled autoregressive processes of order 1. The persistence time for the test time series is estimated from the original records. The similarity function $S(m)$ for these artificial data is estimated N_{sur} times, so that the chosen critical values can be determined., e.g. here the $q_{low} = 5$ and $q_{hi} = 95$ % quantiles of the distribution of similarity estimates. Here, similarity is estimated only for zero lag. Still, it might be advisable to calculate the *lagged* similarity functions if age-uncertain proxy data are investigated and the age uncertainty can not be incorporated through ensemble runs. This is the case for example if dating tables are not available (cf. Ch. 3 and Ch. 5). Then the largest absolute value of the similarity function $S(m\Delta t_{xy})$, around zero lag should be considered for the significance test.

5. Finally, these critical values are used to **threshold** the correlation matrices. If a significant correlation exists between the records i and j, i.e., $S_{est}^{i,j} < S(q_{low})^{i,j}$ or $S_{est}^{i,j} > S(q_{hi})^{i,j}$, the respective entry in the adjacency matrix is set to $A(i,j) = 1$. If no significant similarity is found the entry is set to zero. This is repeated for all similarity estimators and thus n_{sim} adjacency matrices are obtained. The matrices are summed to obtain the final, weighted, adjacency matrix for the network. In this context, the nodes i and j are thus linked, if any $A(i,j) > 0$. Link weight scales between zero (no link) and n_{sim} (all measures find a significant link).

6. The above steps are repeated for each time series ensemble member associated with the node unless age uncertainty can/is to be neglected. **Network measures** can be estimated for each ensemble member and link probabilities are associate with the relative frequency of their occurrence. Visually, strong links in the spatially embedded network indicate strong dependencies. A closer inspection of the time series of these significantly linked nodes could reveal reasons for their association (e.g. archive bias, a common driver phenomenon, causal dependency or teleconnections (c.f. Fig. 2.3 on page 8).

4.1.4 PAN measures

The recent heightened interest in complex networks science has resulted in an abundance of complex networks, or graph-theoretical, *measures*, i.e. statistics that supposedly reflect characteristic node, link or network properties [165, 58, 58]. In the paleoclimate context, the specific challenges require an adaptation and careful evaluation of the commonly employed (climate) network measures, because of age uncertainty and temporal and spatial heterogeneity. In a first step, the validation process for Paleoclimate Networks is therefore naturally restricted to a small subset of available characteristics. Potential paleoclimate network measures considered for the test in the following include the

- the average **link density** (or connectivity),
- the **regional degree** and **cross-link ratio**, and
- shortest path betweenness centrality.

In the following paragraphs I will motivate and derive the basic paleoclimate network measures and give the definitions for the more common complex network measures.

Average link density

For general complex networks, the link density, or connectivity, of a graph G with N_{no} nodes is simply the ratio of realized links amongst the nodes vs. the number of possible links

$$LD'(G) = \frac{\sum_{i,j} A_{i,j}}{(N_{no} - 1)\, N_{no}} \ , \tag{4.5}$$

which is between zero and one.

However, the number of nodes of a paleoclimate network varies, if time-scale uncertainties are large and the average time resolution low. In such cases the minimal overlap amongst the time series is not always fulfilled. Therefore the node number differs amongst the ensemble realizations $N_{no} = N_{no}^l$ and this results in a link density that depends on the realization number l and the link weight LW:

$$LD(l) = \frac{\sum_{i,j} LW_{i,j}^l}{(N_{no}^l - 1)\, N_{no}^l} \ . \tag{4.6}$$

This expression is averaged to obtain the *average link density* for the considered ensemble of time series,

$$LD = \frac{\sum_{l=1}^{N_{ens}} LD(l)}{N_{ens}} \ . \tag{4.7}$$

Cross link probability

Assume network G consists of N_{no}^{G} nodes and N_{ed}^{G} edges. Let us partition this network into nodes in two subnetworks, say, H_1 and H_2 with $N_{no}^{H_1}$ resp. $N_{no}^{H_2}$ nodes each such that $N_{no}^{G} = N_{no}^{H_1} + N_{no}^{H_2}$. Accordingly, the sum of edges is partitioned into the sum of edges within H_1 and H_2, $N_{ed}^{H_1}$ and $N_{ed}^{H_2}$, and edges from H1 to H2, N_{ed}^{1-2}.

The *possible* number of edges if G is fully connected is $N_{ed}^{G} \cdot \left(N_{ed}^{G} - 1\right)$, because self-loops (i.e. auto-correlation) are not considered and the edges are undirected. Similarly, the maximally possible number of cross-edges is equal to the product of the number of nodes in subnetworks H_1 and H_2, $N_{no}^{H_1} \cdot N_{no}^{H_2}$.

The relative frequency of *realized* cross edges,

$$P'_{1-2} = \frac{N_{ed}^{1-2}}{N_{no}^{H_1} \cdot N_{no}^{H_2}} \quad , \tag{4.8}$$

gives the *cross link probability* P_{1-2}

$$P_{1-2} = \frac{\sum_{i \in H_1, j \in H_2} LW_{i,j}}{N_{no}^{H_1} \cdot N_{no}^{H_2}} \tag{4.9}$$

in the case of weighted edges LW with weights between 0 and 1.

Cross link ratio

The cross link ratio $CLR(H_1, H_2, G)$ is given by the cross link probability divided by the overall link probability,

$$CLR(H_1, H_2, G) = \frac{P_{1-2}}{LD} \quad . \tag{4.10}$$

Average and regional node strength

In classical complex network theory, the *degree* D of a node i is a measure of the presumed importance of a node, given by the number of its links to all other nodes $j = 1, \ldots, N_{no};\ j \neq i$,

$$D'(i) = \sum_{i,j \neq i} A(i,j) \quad . \tag{4.11}$$

This classic D' was extended to accommodate for link weights (cf. [116] and references therein), and the notion of the *degree* of a node was replaced by that of a *node strength*, also called *vertex strength* [58, 116]. Using the probabilities of a link as a weight, this gives a *node strength* D^l

$$D^l(i) = \sum_{j \neq i} LW^l(i,j) \tag{4.12}$$

for each ensemble realization l. As the nodes composition of the paleoclimate network changes in time, this node strength is averaged *regionally*, i.e. for nodes within a certain subnetwork, to obtain the necessary independency of the individual nodes in the investigation of regional characteristics.

Shortest Path Betweenness Centrality

Betweenness centrality has been regarded as a measure for local dynamical information flow [116, 8]. The shortest path betweenness of a node k is calculated from the number of *shortest*

paths $\sigma_{ij}(k)$ between all other nodes i and j that pass through node k and their overall multiplicity σ_{ij}:

$$BC'(k) = \sum_{j \neq k} \frac{\sigma_{ij}(k)}{\sigma_{ij}} \quad . \tag{4.13}$$

As the node number in paleoclimate networks is changing over time, is is necessary to standardize the betweenness to obtain a measure that is independent under such changes:

$$BC(k) = \frac{BC'(k)}{(N_{no} - 1)(N_{no} - 2)} \quad . \tag{4.14}$$

4.2 KIMONO: A semi-empirical Asian Summer Monsoon model

4.2.1 Asian Monsoon Dynamics: A very brief overview

In the following I will give a brief introduction into the monsoon as a climate phenomenon in the global and regional context. For an exhaustive discussion of the global aspects of monsoon I refer the reader to the book on "The global monsoon system" [28]. An in-depth and very detailed discussion of "The Asian Monsoon" is found in [172].

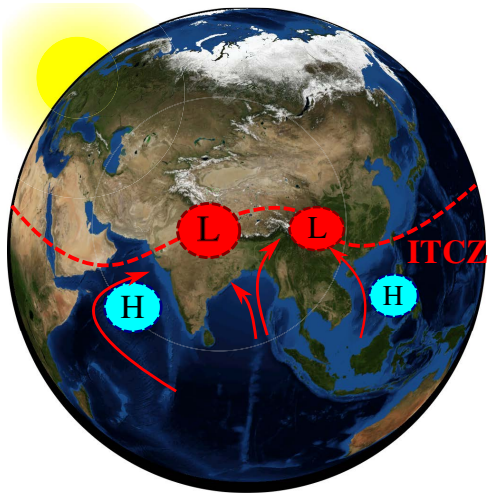


Figure 4.4: Schematic illustration of the Asian summer monsoon circulation. Fundamentally, its circulation is driven by seasonal insolation changes, resulting in differential warming of land and oceans and northward migration of the ITCZ and the corresponding balancing winds.

The word *monsoon* originates from the arabic word مَوسِم *mausim*, meaning *weather* or *season*, and relates to the observed seasonal reversal of surface winds, accompanied by a strongly seasonal wind and precipitation pattern that changes as the Inner Tropical Convergence Zone (ITCZ)

moves northwards during Northern Hemispheric (NH) summer. The ITCZ deforms primarily in response to the thermal gradient between landmasses and the surrounding oceans, thus bending the global Westerly trade winds [172]. In the NH, this means a wet summer vs. a dry winter in most parts. Some regions can receive rainfall both when the ITCZ moves northwards in summer and when it moves south in Winter. For Sri Lanka, for example, the monsoonal winds pick up moisture south of the island in summer, and when the ITCZ moves south in autumn, the water vapor source along the pathway is found North of Sri Lanka.

Several regional *monsoon systems* have been identified, ranging from the South Asian, the East Asian, Australian, African to the Mexican/North American and South American monsoon systems, spatially separated but intimately linked through the ITCZ and the annual insolation cycle. Fundamentally, as illustrated in Fig. 4.4, the seasonal irradiation changes lead to thermal imbalances between the respective land masses and adjoining oceans, resulting in pressure gradients that are balanced by the surface winds.

The Asian monsoon system comprises the Indian and East Asian subsystems. Spanning from the Arabian Sea all the way to East Asia, and from North Australia into Central Asia it is a major player in the global climate system. In the sub-tropical regions of Asia, the summer season and thus the *Asian Summer Monsoon*(ASM) is associated with the majority of annual precipitation. Roughly 60% of the world's population live in this region and depend crucially on the summer monsoon's delivery of moisture, with droughts and floods associated with starvation and dwindling economies [82, 34, 24]. ASM dynamics are determined by the interplay between the Indian Summer Monsoon (ISM) and the East Asian Summer Monsoon (EASM). The predominant regions of influence, and main wind directions, of ISM, EASM and the continental Westerlies are given in Fig. 4.5. They can be roughly divided at 100° Eastern longitude [177]. The dependency – and interplay – of the subsystems on each other is a topic of ongoing research [177, 187, 26] and will be discussed in more detail in Ch. 5.

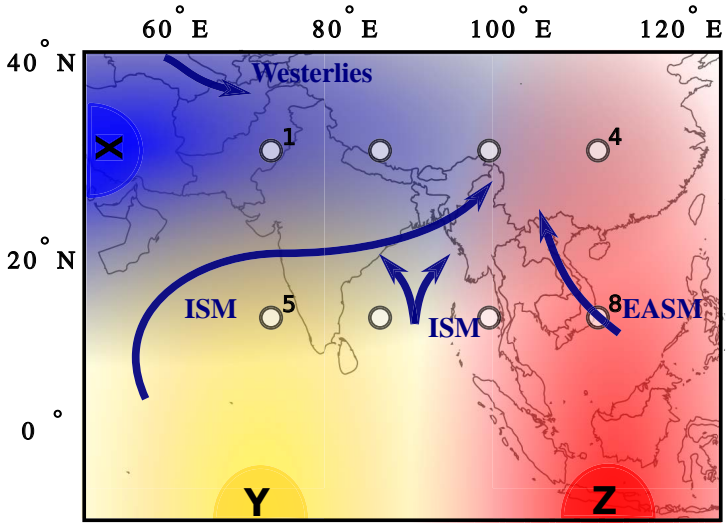


Figure 4.5: Map showing the main wind directions of the Indian and East-Asian summer monsoon systems. The three inter-regional inflow corridors are modeled as sources of variability in the semi-empirical monsoon model KIMONO: The Indian Summer Monsoon (ISM) with a longitudinal (X, in blue) and a latitudinal component (Y) and the East Asian Summer Monsoon (EASM) with a latitudinal component. The dynamics in the model, governed by the respective strengths of the source flows, are sampled at the node locations (black dots).

4.2.2 KIMONO: model philosophy

KIMONO is designed to be a simple semi-empirical model of information flow in a spatially extended domain. It is a statistical toy model with three independent spatial components that react differently to applied external forcing. Using this simplified model, I want to investigate how changes in the external forcing parameters are reflected in a (paleo)climate network topology and the introduced network measures and to what extent they are affected by the underlying spatial topology (i.e. observations on a regular grid or heterogeneously distributed nodes).

Let us assume that we can view the Asian Monsoon domain in summer as a region with three main wind systems (cf. Fig. 4.5) which each extend only zonally, i.e. laterally *or* longitudinally. While the true wind pattern might be significantly more complex, I argue that this can be viewed as a statistical decomposition of the mean summer surface wind field into simplified lateral/longitudinal components $V_X(lat, lon)$, $V_Y(lat, lon)$ and $V_Z(lat, lon)$. These fields are assumed to be Gaussian-modulated unidirectional fronts with a velocity at position $\vec{p}$ and time point t

$$v_{A,t}(\vec{p}, m_A(t,T), w) = m_A(t,T)e^{-(p_x - p_{0,x})^2/W} \quad , \tag{4.15}$$

with a spatial half-width W. The maximal amplitude of the velocity, $m_A(t,T)$,

$$m_A(t,T) \quad = \quad B_A + \alpha T \tag{4.16}$$

$$\tag{4.17}$$

is found in the center of the Gaussian front, as in Fig. 4.5. Here, B_A is the baseline strength of the component's flow, and α is its amplitude, or susceptibility to the external forcing, represented by the parameter T, $T = [-1, \ldots, 1]$. The velocities for sources Y and Z are defined analogously, and the chosen values are given in Table 4.1.

Each of these fields originates from a source at a position $\vec{p}_{src}$, and each of the sources X Y and Z is associated with a "climate process", X_t, Y_t and Z_t which represents the annual mean of a hypothetical climate variable in the year t. Thus, the model is restricted to modeling inter-annual variability.

The amount of dynamical information about the climatic process at the source that flows along one of these fields, say, from source X, to a point at a position $\vec{p}$ in its region of influence is approximated by a *variance factor* $f_A(\vec{p}, T)$. By construction, the square of the variance factor is proportional to the amount of variance shared between the source A and the time series at point $\vec{p}$: $f_A^2(\vec{p}, T) \propto \sigma_A(\vec{p})$. This factor is determined as the height of a hypothetical temperature peak time series that is inserted and observed in the flow X at the source origin relative to the height observed at a position $\vec{p}$ after a given time later. $f_A(\vec{p}, T)$, $f_B(\vec{p}, T)$ and $f_C(\vec{p}, T)$ are calculated using an approximation of the Advection-Diffusion equation which governs the transfer of physical quantities (i.e. heat, particles or energy) in a physical system due to diffusion and/or large scale motions of fluids in the system [153]. The derivation of these factors and the solution of the Advection-Diffusion equation for Gaussian-modulated unidirectional flows is given in Appendix 1.

The factors depend on the observation position p, source positions p_A, p_B and p_C and field velocity, $v_A(t,T)$:

$$f^*_{A,t}(\vec{p}, T) = f(\vec{p}_A, \vec{p}, v_{A,t}, c, s) \quad . \tag{4.18}$$

At each point in space, and for each time point t the factor sum $g(t,T)$,

$$g(t) = f^*_{A,t}(T) + f^*_{B,t}(T) + f^*_{C,t}(T) + f^*_{D,t} \quad , \tag{4.19}$$

is used to standardize the factors, i.e. $f_{A,t}(T) = \frac{f^*_{A,t}(T)}{g(t)}$. The factors depend implicitly on each

Table 4.1: KIMONO source and flow attributes.

	Source	Position	Strength B	Ampl. A	Width W
ISM *(long. component)*	X	$(30, 55)$	85	$\alpha = 70$	1200
ISM *(lat. component)*	Y	$(-15, 70)$	10	$\beta = -5$	100
EASM	Z	$(-15, 112.5)$	30	$\gamma = -12$	100
White noise	N	everywhere	—	const.	—

other due to the normalization process of equ. 4.19. The processes X_t , Y_t and Z_t are uncoupled AR(1) processes of unit variance, and with a persistence time of $\tau = 6.4$ years (i.e. $\Phi = 0.7$, please refer to Section 2.5.3 in Chapter 2 for more details). This value for τ, resp. Φ was chosen as it is compatible with the order of magnitude estimated for ASM paleoclimate proxy records [128].

At point p in the region of interest, the local climate "history" $R_t(\vec{p})$ is then assembled from the contributions from the three components at this point, $f_{A,t}(\vec{p}, T)$, $f_{B,t}(\vec{p}, T)$ and $f_{C,t}(\vec{p}, T)$ and a constant term $f_{D,t}(\vec{p})$ describing the amount of exclusively local uncorrelated and Gaussian-distributed observation noise with zero mean and unit variance, $N_t = N(0, 1)$:

$$R_t(\vec{p}) = f_{A,t}(\vec{p}, T)X_t + f_{B,t}(\vec{p}, T)Y_t + f_{C,t}(\vec{p}, T)Z_t + f_{D,t}(\vec{p}, T)N_t \quad . \tag{4.20}$$

4.2.3 Model setup

Table 4.2: Spatial setup for the KIMONO model experiments

	Grid	Heterogeneous
Number of nodes	42	36
Regional span [lat]	$[-10; 40]$	$[-7; 39.5]$
Regional span [lon]	$[60; 120]$	$[66; 115.5]$
Node distribution	regular grid	paleoclimate record origins, c.f. p. 5.1

KIMONO generates synthetic time series at locations in its integration domain. The interdependencies amongst the time series depend on their relative position in the considered flows and the given forcing. This way, it is possible, for example, to test whether two observed proxy time series can actually be linked through such convective flows and diffusion, how changes in the spatial interdependency structure, the network, are reflected in the network measures, and how the estimates change if the spatial node distribution is changed. Although each time series is distinct, as the influence of the different components is location-dependent, time series located in close proximity are similar, and the amount of variance shared with the components' sources is both location- and forcing-dependent. In this Chapter, transient forcing model runs are performed for two spatial sampling types, a grid and a dataset of locations of paleoclimate records (c.f. Table 4.2 and Table 5.1) throughout the ASM domain. The regional span of the spatial sampling and the node numbers are comparable, although the archive locations are spaced closer at the center of the ASM domain, while the grid also samples the areas over the Indian Ocean,

as shown in Fig. 4.6. Please note that the temporal sampling of the records is not relevant here, the time series considered are all simulated for unit (annual) intervals, and for a length of 200 years. Networks are computed for 20 time slices of a 4000 year transient simulation, during

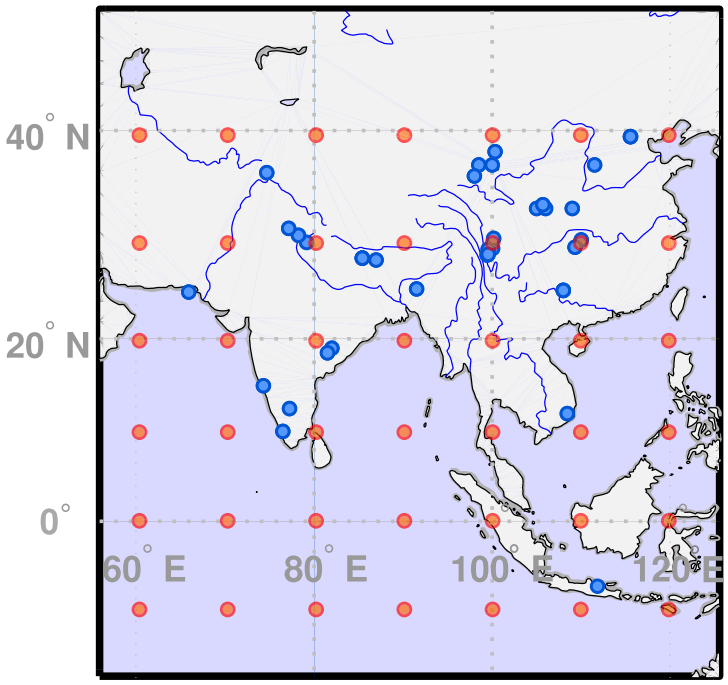


Figure 4.6: Spatial heterogeneity is the central challenge in paleoclimate dynamics reconstruction. To evaluate how the different network measures are effected by this feature, KIMONO model results are compared for a gridded test case (red dots) and for the positions of suitable paleoclimate archives (blue dots), c.f. also Table 5.1 in Ch. 5.

which the model forcing parameter, T, was increased consistently from its minimum value, -1, to 1. To ensure the robustness of the intended spatial inference against estimation errors for these relatively short time series, 100 transient realizations of the time series were analyzed separately. The effective forcing for nodes in the different regions is given in Fig. 4.7 – with increasing forcing , the ISM component variance fraction increases, while the other factors decrease. As the time series are all consistently sampled, time- and resource-intensive surrogate time series are not generated. Instead, the network is thresholded such that the q_{thresh} percentage of *strongest* links are considered significant. Note that this thresholding has to take place at the level of the similarity matrices for each measure, as the scales of variation for them might be different.

4.3 Validation of PAN using KIMONO

In this section I first investigate the topology of the networks that are reconstructed for different forcing parameters, "ISM off", equivalent to forcing $T = -1$, "Coexistence" with $T = 0$ and "ISM on" with $T = 1$, focusing on the difference resulting from the underlying node topology (regular or scattered).

Then I show how changes in the topology of the networks are reflected in the network *measures*, and how robust they are with respect to spatial sampling.

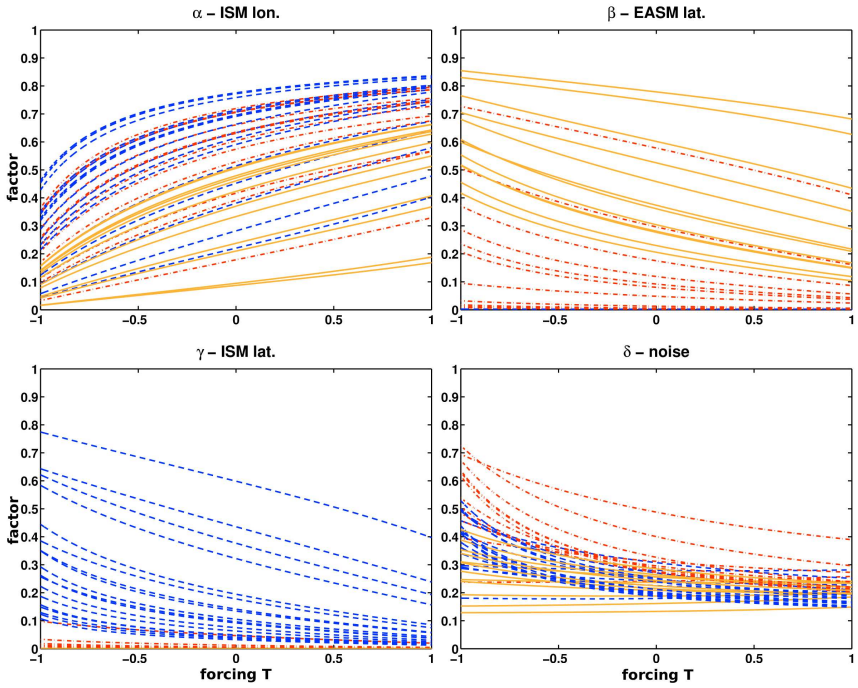


Figure 4.7: KIMONO forcing: Contribution of the different variance factors at each grid-point and forcing T. The change in variance depends on the grid point location. The curves for ISM (longitudinal) nodes are the dashed blue lines, the latitudinal ISM component in yellow and those for the EASM zone are given in the dashed-dotted red lines. While the longitudinal ISM component contribution (top left) increases for all locations, the latitudinal components and the noise factor δ in the bottom right decrease with increasing ISM dominance.

4.3.1 Topology of the observed networks

Varying the forcing parameter T in the KIMONO model from its minimal to its maximal value, i.e. gradually increasing it from -1 to 1, the maximal velocity in the Advection-Diffusion Equation (Eq. 2 in Appendix 1) is modulated. Since here the amount of variance transported along a flow is proportional to the velocity, the synchronizing reach of the three "wind components" change, as illustrated schematically in Figures 4.8a – 4.8c with changing forcing. The components are tuned to react proportionally to the effected forcing, but compete at each point in space. Therefore, the fraction of variance explained by the components in each location changes in a nonlinear, nontrivial way (c.f. 4.7). As the reach of the ISM component increases, the other components lose relevance.

In the following I investigate the reconstructed network for the three maximally different stages,

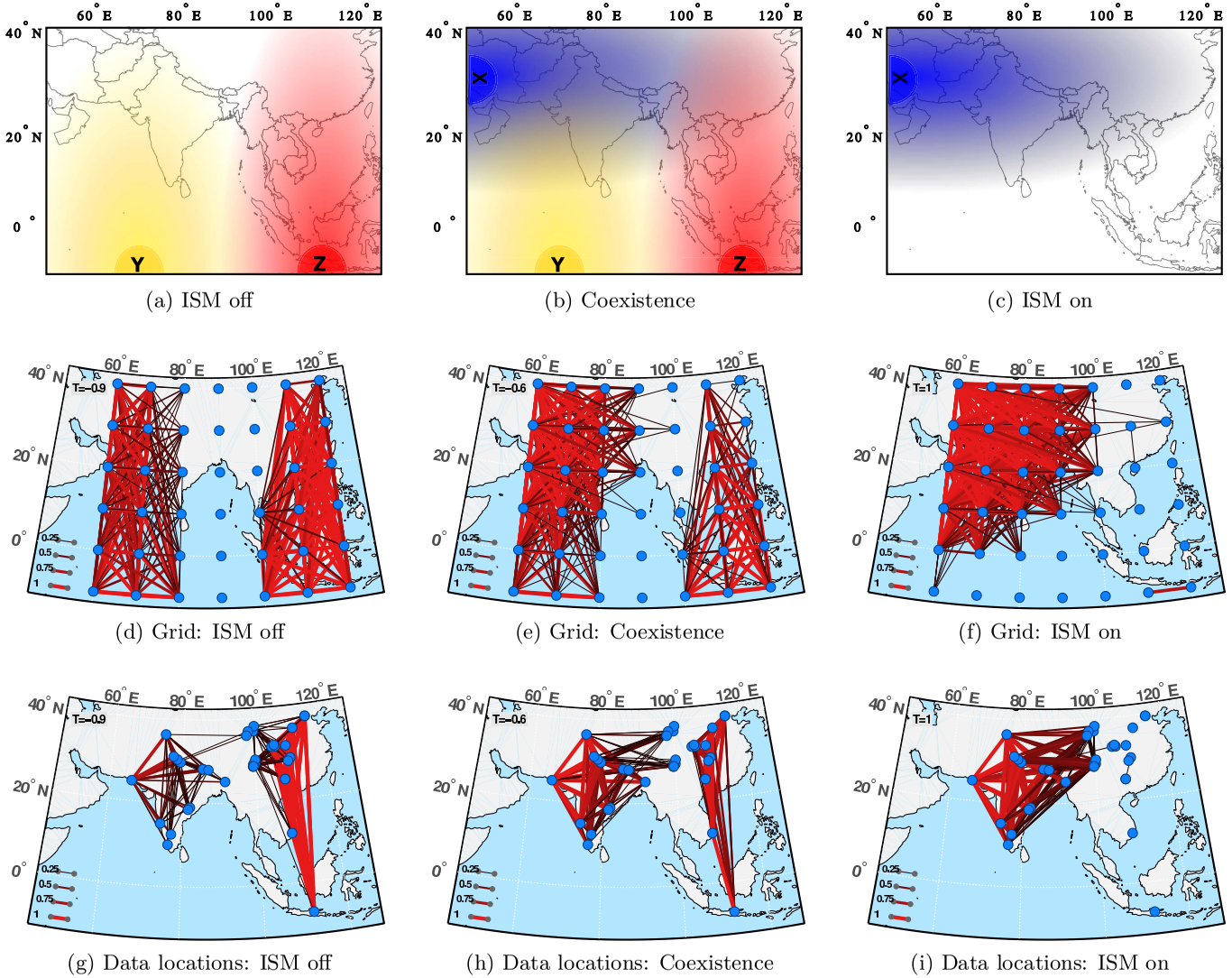
1. *"ISM off"* with minimal forcing, corresponding to $T = -1$,
2. *"Coexistence"* corresponding to $T = -0.65$ and
3. *"ISM on"* with maximal forcing corresponding to $T = 1$.

Topology of the network sampled on the observation grid

1. Fig. 4.8d, *"ISM off"*: The two vertical components of EASM and ISM dominate the local variance splits. This is well reflected in the reconstructed grid-based network, which shows a clear separation of the two components. The ISM component covers the longitudes 60-80°E, while the EASM component covers the longitudes 100-120°E, and no strong links appear between the two, because the central nodes do not lie in a synchronizing region.
2. Fig. 4.8e, *"Coexistence"*: With increased forcing, the longitudinal ISM component strengthens, and the latitudinal components lose in relation, as they have opposite sensitivities to forcing (c.f. Table 4.1). The reconstructed network is still split, as the strongest links are in the core of the regions of influence. However, strong links, originating far in the West, connect to the 90°E grid-points, and weaker links extend even beyond. The EASM half of the network has retracted southward, in agreement with the decreased relative forcing strength.
3. Fig. 4.8f, *"ISM on"*: At maximal forcing, the ISM component dominates the network, which now has one single core region.

Topology of the network sampled on the data locations

1. Fig. 4.8g *"ISM off"*: While there is a clear concentration of the strong links in the regions of 60-80°E and 100-120°E, there exist a relevant number of links with low weight that connect the two regions. These links occur because the link density is fixed at 20% for these model runs.
2. Fig. 4.8h *"Coexistence"*: With increased longitudinal component, the ISM region is fully connected and strong links extend up to 90°E. The EASM core region shows few strong links but is still well-separated from the ISM component.
3. Fig. 4.8i *"ISM on"*: As the ISM component profits from its strong forcing the EASM component vanishes, and only few links with low occurrence and weight remain.



(a) ISM off (b) Coexistence (c) ISM on

(d) Grid: ISM off (e) Grid: Coexistence (f) Grid: ISM on

(g) Data locations: ISM off (h) Data locations: Coexistence (i) Data locations: ISM on

Figure 4.8: Extreme points of the modeled Asian Monsoon dynamics: Schematic illustration of input variance factors (a)-(c) and networks reconstructed on grid (d)-(f) as well as the actual data positions (g)-(i).

Comparison of the reconstructed topologies

The changing strength of the synchronizing components due to the varied forcing is reflected by the extent of the synchronized regions in which the grid point time series show the strongest similarities. Although differences exist, the networks reconstructed from grid and heterogeneous locations bear large resemblance and relate to the underlying forcing structure. The spurious links in the data-based "ISM off" scenario are due to the use of a specific and rather high link density of 20% and disappear if it is lowered.

4.3.2 Validation of the network measures

Network measures are statistical estimators that reflect properties of individual nodes as well as regional or global characteristics of the network. The spatio-temporal changes, visible as the KIMONO model is driven with different forcing, should effect consistent transitions in measures suitable for the investigation of spatio-temporal changes. Furthermore, these transitions should also be detectable under varying node distributions, i.e. for a grid-structure as well as for heterogeneous node distribution. I computed network measure expectation values from 100 transient KIMONO forcing runs. The gradually increased forcing results in a transition from a bimodal, laterally synchronized network with separated components to a widely connected state. Figure 4.9 gives an overview over the obtained results for the network measures.

Cross link ratio

Grid: The *cross link ratio* (Fig. 4.9a), calculated as the ratio of cross link probability over overall link probability, increases monotonously for the grid-based networks and saturates well before maximal forcing is reached. It crosses the threshold of equal probabilities, transitioning from values of .6 to 1.2. This means that while in the early, bimodal, stage the cross link density is about 60% of the average link density, it reaches 120% when the synchronized region spans the whole network. Thus, at low forcing the cross link density is significantly *lower* than the overall link density, with the network being effectively partitioned. At high forcing the co-varying region spans across the former separate parts, resulting in a *higher* cross link density than average link density.

Data locations: The expected path of transition for the heterogeneous network differs from that observed for the grid topology by a negative offset, but parallels it otherwise.

Average node strength in the intermediate domain

The *intermediate domain* (85-105°E) between the defined *model* ISM/EASM core regions is not synchronized by an external source of variability at the outset of the transition experiments. Thus, nodes within this domain are expected to have no, or few, connections, resulting in weak node strength. This should change, however, as the longitudinal ISM component strengthens with increasing forcing, when these nodes fall into its region of influence. Please note that this does not tie exactly with the actual ISM/EASM transition region, which is situated 10-15° further to the East. This adaptation was necessary because the implemented flow paths in KIMONO are modeled for simplicity either lateral *or* longitudinal, and the core interaction region around 105°E/30°E should be reachable for both flows.

Grid: Indeed, as shown in Fig. 4.9b the average node strength in the intermediate domain rises from a low level to full connectivity (node strength equal to $N_{no} - 1$). After an initial short decline it rises first rapidly and then more slowly.

Data locations: The overall pattern of a short decline, a quick rise followed by tapering off at a maximal level, is similar for grid and heterogeneous locations and here, too, full connectivity is reached. Still, the amplitude of the change is lower for the heterogeneous locations.

Regional node strength ratio

The regional node strength ratio is designed to highlight potentially differing network properties in subregions of the network. In the ASM context, it is computed using Eq. 4.11 as the ratio of the average node strengths in India vs. that in China. If both subnetworks are equally well connected within the overall network, their average node strength should be similar, and the node strength ratio should equal unity, as indicated by the gray line in Fig. 4.9c.

Grid: Starting off at equally well-connected subdomains, the Western, Indian part of the network gains importance with increased forcing and the node strength ratio settles after a short growth slightly below 4. Thus, nodes in "India" are associated with four times the link weight when compared to those in "China".

Data locations: For the heterogeneous sampling scheme the line of equal node strength is crossed *later* than for the gridded data. The following incline, contrastingly, is sharp and the plateau reached gives eightfold node strength for India vs. China, accompanied by a large uncertainty in this estimate. This is consistent with the stronger representation of "China" in the used grid, c.f. Fig. 4.9f.

Shortest Path Betweenness

Shortest Path Betweenness centrality is a measure developed to infer the presumed relative importance of nodes and regions [58, 116, 8, 40]. Furthermore, Donges et al. [41] found that "betweenness centrality allows to measure the importance of localized regions on the earth's surface for the transport of dynamical information within a climatological field in the long term mean", and stated that "information is transported by advective processes, where the assumption of information traveling on shortest paths can be substantiated by extremalization principles". As such it should, in principle, also be an interesting measure for paleoclimate network applications.

Still, as a node-based measure it is not possible to compare it directly when using different spatial sampling schemes. The characteristic dynamical changes should be reflected in regional properties. Based on this presumption, betweenness centrality estimates for nodes in the intermediate and central region (85-105°E) are averaged to obtain a domain estimate. The results, shown in Fig. 4.9d, are inconsistent for the first segment of the transition experiment, in which the largest dynamical changes occur. On the one side, the betweenness estimates for the grid *in*crease slowly, albeit with a comparatively large uncertainty. On the other hand, the estimates for the heterogeneous locations *de*crease initially from large initial values to then stay on a flat plateau.

4.3.3 Results for the network measures

Cross link ratio: While the paths for different node distribution differ substantially in their amplitude, baseline and speed of increase, both show the overall increase in subdomain-connecting cross-links. The initial decline observed for the data topology results from spurious cross-links, as the overall similarity level, see Fig. 4.9e, is much lower than at higher forcing levels. These cross-links appear randomly at low forcing, but disappear if the link density is chosen more conservatively.

Average node strength in the intermediate domain Heterogeneous sampling apparently does not strongly affect the node strength in the intermediate domain. Both sampling schemes reflect the transition of these nodes from being irrelevant, to heavily tied to the rest of the network.

Regional node strength ratio The expected transition path for the relative average strength of nodes in India vs. that of those in China is significantly different for differing sampling schemes. The general feature, however, that repeats itself is that proximity to the source clearly results in higher node strength. There is nevertheless large uncertainty associated with the transition path for the data-based network.

Betweenness: The betweenness centrality estimates are inconsistent for the different sampling schemes, and the error margins spread widely. The counterintuitive initial decline observed for the cross-link ratio parallel those observed in the cross-link ratio and average node strength, and likely result from spurious cross-links.

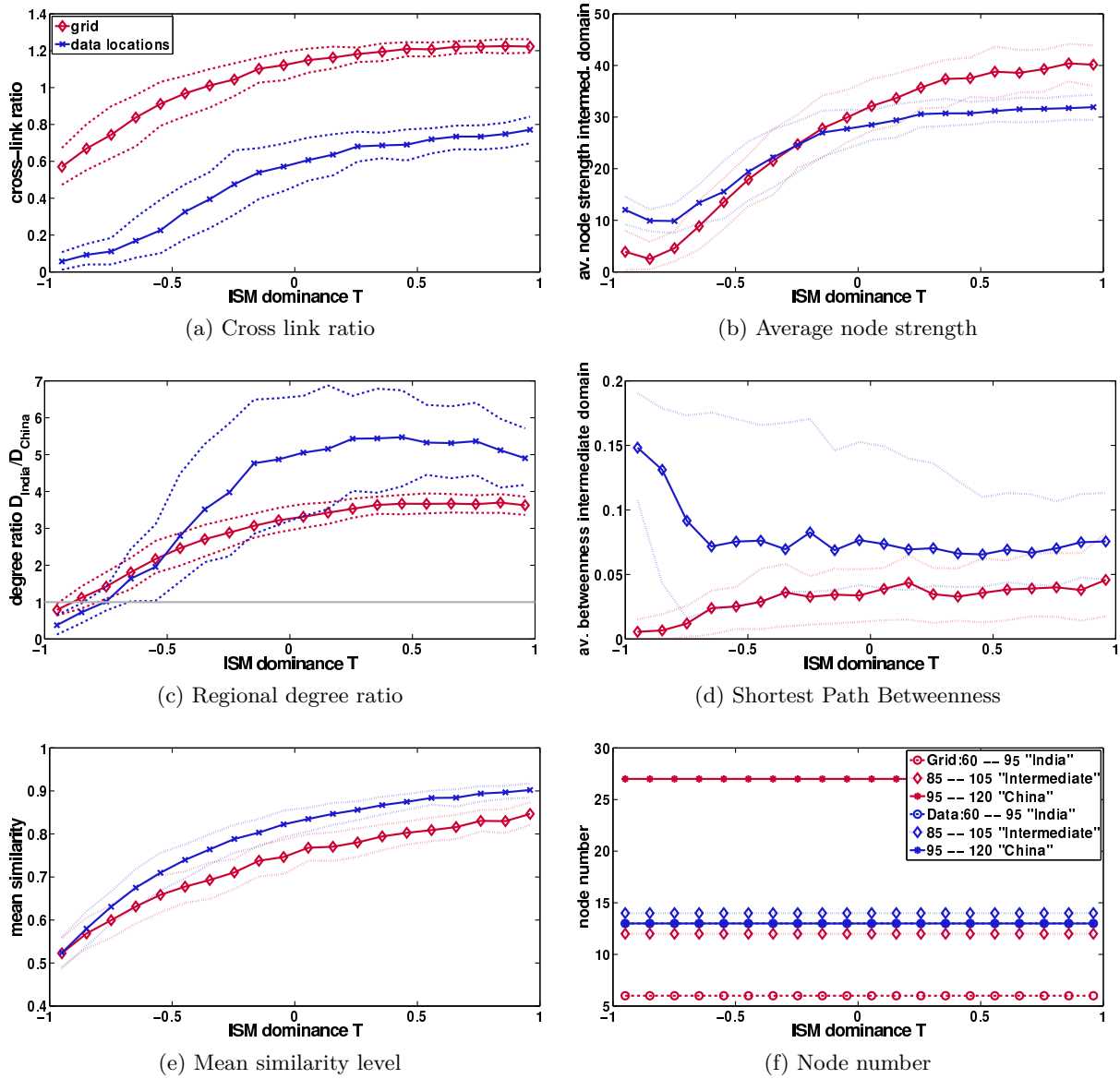
4.4 Discussion of regional changes and inter-regional information flow

In the following I discuss the potential of the paleoclimate network approach for the reconstruction of paleoclimate dynamics, with a focus on the effects of spatial heterogeneity.

Three major regions are relevant for KIMONO dynamics: The ISM region, the intermediate region and the EASM region. While the first and the latter start of as independent subdomains of the network, they are connected by the longitudinal ISM component at increased forcing. The distinct dynamical features include (i) bimodality vs. later unimodality (ii) increasing size of the spatially synchronized region and (iii) increased flow of dynamical information through the increasing strength of the longitudinal ISM component.

The initial *bimodality* of the network is directly visible in the network, and in the low cross link ratio and average node strength in the intermediate domain. The developing connection between the two parts is visible in the network reconstruction and in cross-link ratio, and the intermediate domain node strength. Although the absolute values of the between cross-link ratio are not on the same scale as that for the grid, a significant increase occurs. Thus, the change in model dynamics can be inferred if the node topology does not change and if the sampling bias can be quantified. Node strength in the transitional region shows smaller sampling dependent deviations and could thus be a more robust measure of the importance of intermediate regions. The regional degree ratio shows, however, that spatially biased sampling can have large effects: As only one node of the paleoclimate network samples close to the EASM source, and most of the rest clusters far North of it, the synchronizing influence of this source quickly disappears as the ISM region grows. Located at the fringe of the ISM component, these nodes also get less ISM input than all its westwards neighbors. This results in few links and an under-representation of the Chinese part of the network, caused by a combination of model shortcomings and data sparsity. Such effects have to be addressed before a comparison of networks with changing node architecture is possible. The increasing strength of the ISM component is directly visible by its growth in the reconstructed network, and additionally reflected in the increasing cross-link ratio. The source-region of the modeled pathway is robustly characterized by higher node degree. Increasing information flow from the ISM to the EASM core region is indecipherable using Shortest path betweenness, which is sensitive to sampling changes and spurious links. Nevertheless, provided the potentially changing node topology is addressed, the cross-link ratio could be a sufficient measure to quantify changes in inter-regional dependencies.

The previous sections were concerned with heterogeneously distributed nodes in a spatially embedded network. Still, what remains to be investigated is the influence that varying node



(a) Cross link ratio

(b) Average node strength

(c) Regional degree ratio

(d) Shortest Path Betweenness

(e) Mean similarity level

(f) Node number

Figure 4.9: Most network measures reflect the increasing change in network structure occurring with changing ISM dominance in KIMONO (a)-(f), but they show different sensitivities. The spatial distribution of node positions influences the absolute value of the network measure, but the general trends are consistent for the underlying grid and the heterogeneous data positions.

numbers have on estimated network measures. This could be done by simulating constant KIMONO 'climate dynamics' and removing nodes iteratively. The relevance of the spatial position of a node can then be assessed by the discrepancy between the network measures estimated with/without it in comparison to the expected value for high-resolution sampling.

4.5 Summary

The key questions addressed in the beginning of this chapter concerned (i) the ability to reconstruct the ASM model dynamics using the network approach, (ii) the prospects of inducing regional changes, (iii) the potential to extract inter-regional information flow and finally (iv) the effect of spatial heterogeneity on the analysis.

(i+ii) It is possible to reconstruct the spatio-temporal changes in the semi-empirical ASM model, also for varying node topology, though the results are limited to qualitative statements if the node topology changes. Model dynamics are reflected in the reconstructed network, specifically its link strength distribution, and in network measures such as regional average node strength, and node strength in the transitional zone, the intermediate domain outside the latitudinal source influence.

(iii) Inter-regional information flow, or, in the context of the KIMONO model, spatial distribution of variance, cannot be inferred using Shortest Path Betweenness, as it is found to be too sensitive to irregularities, and no clear dynamical signature can be found in the transition experiments. The cross link ratio is a better alternative, though sampling biases have to be taken into account for its analysis.

(iv) Spatial heterogeneity has strong effects, both on the reconstructed network and on the network measures. It manifests itself in

- biases in network measures, that can be negative (cross-link ratio) as well as positive (regional degree ratio),
- increased variance in the estimates (Betweenness Centrality, Regional Degree ratio),
- and the amplification of effects due to node clustering (Regional degree, reconstructed network)

If networks with varying spatial sampling are investigated, care has to be taken to perform adequate significance tests, to ensure that spurious sampling effects can be distinguished from real climate processes.

The developed ASM model KIMONO is a toy model that can not be expected to reflect actual monsoon dynamics. In reality, local, global and external forcing influences local climate processes. In the KIMONO model world, information transfer and local climate processes are governed solely by physical flows. Processes external to the ASM domain are not considered but can, in nature, lead to increased correlation in the whole or parts of the ASM region. One of the desired features to improve realism, for example, would be the inclusion of regional sources of variance, i.e. by region-dependent noise terms. Then the propagation of information could be considered serially, and causality-sensitive, directed measures (Granger causality, ESF) could be tested. In the model simplicity, however, also lies its strength, because it is possible to interpret the results with respect to its dynamics, a task that is infinitely more complicated if such "pseudo-proxy experiments" are conducted with actual global climate models (GCMs) [148, 171, 97]. Unlike GCMs, KIMONO is

computationally efficient, therefore large ensembles of time series for pseudo-proxy experiments can be generated. Using KIMONO, hypotheses concerning local vs. global drivers of local ASM climate dynamics can be tested specifically, because it models the propagation of local climate variability through convection and diffusion. For large-scale dynamical and coupled GCMs with their multitude of output variables and parameters, cause and effect are more difficult to discern. Thus, it provides an excellent opportunity to assess whether and how spatio-temporal dynamics of a given paleoclimate dataset are affected by age uncertainty and spatio-temporal heterogeneity and sparsity.

5 Testing temperature-modulated dependency in the Asian Summer Monsoon dynamics of the last millennium

In this chapter I use the developed Paleoclimate network approach to analyze Asian Paleo-Monsoon dynamics[1]. The challenges inherent in such a task are manifold: the sparsity and heterogeneity of the underlying paleoclimate records, their chronological uncertainties and their archive plurality. However, the Asian Monsoon is an important component of global climate, and its driving mechanisms, as well as teleconnections to other climate subsystems, are far from being well-understood [175, 26, 187, 89]. If it is therefore possible to shed light on its evolution using complementary approaches to those employed routinely it is no small contribution.

Provided the average temporal resolution of multiple climate records is comparable, irregular sampling in time can be addressed by adequate similarity measures, as developed and presented in Chapter 2. Chronological uncertainties, can be incorporated directly using ensembles of depth-age relationships, as shown in Chapter 3. Spatial sampling effects are to be considered, but can be probed within the PAN approach using the toy model for ASM dynamics (Chapter 4).

The key questions addressed in this Chapter are the following:

- Can I discern distinct temperature-change induced spatio-temporal patterns for the Asian Monsoon domain in the last millennium, using paleoclimate networks and the semi-empirical ASM model?

- Sub-decadally resolved paleoclimate records from the Asian Monsoon domain are sparse, and few and fewer exist the farther back in time the analysis is extended. Do the effects of spatial and temporal sparsity and chronological uncertainties preclude paleoclimate network analysis?

- Paleoclimate archives may record different climatic parameters, e.g. temperature, *or* precipitation, *or* both. Reconstructing *information flow* instead of physical parameter dependencies, common variability is extracted from heterogeneous archives. Do we observe archive-dependent bias effects, e.g. are tree records more likely to be similar to other tree records, than to stalagmite records? In other words: is the dynamical *information* they convey systematically different?

The methods are ready, it is time to let the data speak for itself.
After a brief literature review focusing on the Asian Monsoon, its evolution in the recent past and its driving mechanisms, I will shortly outline the methodological background of the analysis. Subsequently, the dataset compiled for this study will be introduced and the required preprocessing steps described. The results and discussion sections revolve around the key questions identified above, before I can come to their, to some extent, final answers.

[1] This Chapter is partly based on my previously published paper on "Late Holocene Asian summer monsoon dynamics from small, but complex networks of paleoclimate data" [129].

5.1 Introduction

5.1.1 Asian Summer monsoon dynamics

Monsoonal precipitation dynamics and their possible change due to global warming are a matter of political and public concern in most of South-East Asia, and especially in India and China, as lives and prosperity depend critically on the monsoons' rainfall delivery [34, 82, 189]. The Asian (Summer) Monsoon has shown abrupt changes in the past and its intensification (weakening) was likely concurrent with cultural prosperity (demise) [24, 25, 190, 122, 56]. The Asian monsoon system is comprised of two main sub-systems, the Indian Summer Monsoon (ISM) and the East Asian Summer Monsoon (EASM) (Fig. 5.1), both mainly driven by seasonal changes in the land-sea thermal contrast and related atmospheric pressure changes.

The Intertropical Convergence Zone (ITCZ) plays a governing role in monsoonal circulation and variations of its mean northward extent have been linked with summer monsoon strength [20, 53, 88, 146]. The defining geography (composition of landmass, mean altitude, position and extent of surrounding seas) however, is quite different for ISM and EASM. The extent to which the two sub-systems interacted in the past is a matter of current research [26, 87, 177, 189, 191, 31]. As a third player, the mid-latitude westerlies dominate the area north and west of the (variable) monsoon boundary [31]. The relative strength of these circulation systems and thus their areas of influence, varied in the past [65, 102, 177], and our knowledge about the complex spatio-temporal processes and variability behind them is insufficient [34].

Numerous paleoclimatological studies focused on the reconstruction of individual climatic parameters, such as moisture or precipitation [17, 94, 118, 124, 143, 177, 189, 186], temperature [186], or droughts [17, 34, 146, 186] by use of proxy records. Furthermore, linkages among the Asian Monsoon system and the North Atlantic realm [62, 69, 88, 178, 175], El Niño/ Southern Oscillation (ENSO) [142], and solar forcing [61, 176, 190] have been explored. However, the mechanism(s) and variability of the interactions between ISM and EASM during the Holocene (and beyond) remain far from being fully understood [175, 189, 177]. Using numerical meta-analysis and reconstructions of moisture indices, Wang et al. proposed that the evolution of ISM and EASM for the Holocene was asynchronous on centennial timescales [177]. However, the spatial distribution of the paleoclimatic records used in the study of Wang et al. did include only four records from India (out of a total 92) and focused mainly on China and Tibet, with no record in the ISM domain below 27°N [177]. It is important to note that the currently general low number of datasets from the Indian peninsula might lead to systematic biases towards the Tibetan plateau and China, complicating or even precluding meaningful interpretation of results, a caveat that must be accounted for.

Based on ensemble runs of a coupled climate model run with anthropogenic forcing, May found an increase in monsoonal rainfall, accompanied by a decrease in the intensity of the overall lower-tropospheric large-scale circulation at a warming of 2°C relative to pre-industrial ISM conditions [101]. Derived from global climate modeling results and observations, an overall stagnation in precipitation but a redistribution towards extremes (prolonged dry and wet spells) was supported in [82]. Decreasing reliability of rainfall and increased variability of precipitation amounts would have disastrous impacts on rain-fed agriculture all over Asia.

In the paleoclimatic context, it can be asked whether the weakening of the large-scale circulation associated with a warming scenario, as found for the time period 2020–2200 AD in the modeling study by May [101], is paralleled by an increased influence of the ISM on the EASM domain during the MWP (1100–700 yrs BP) and during the recent warm period (RWP, 1850–1980 AD), in contrast to a potentially diminished influence during the LIA (100–400 years BP). Given that the Asian Summer Monsoon is, amongst other factors, differential-heating driven, and thus modulated, to some extent, by northern hemisphere temperature, it is reasonable to

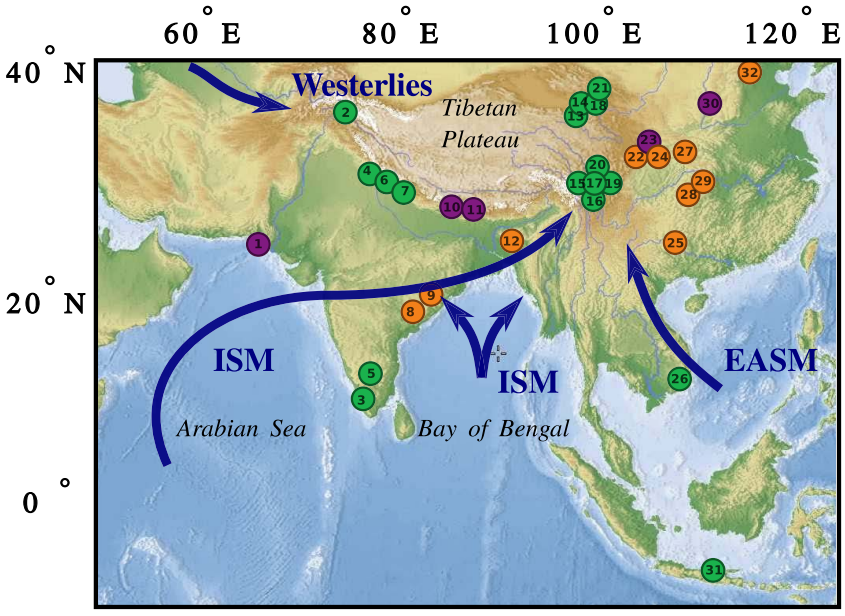


Figure 5.1: Study area with generalized summer wind directions of the ISM and EASM, the westerlies (blue arrows), as well as the spatial coverage of the records considered in the paleoclimate networks. Node numbers were assigned according to the respective site longitude and furthermore refer to the entries in Table 5.1. Sites that are at close proximity show slightly displaced to prevent overlap of dots and labels. Colors indicate the type of archive: *green* dots stand for tree sites, *orange* for cave sites and *purple* ones for other archives.

hypothesize that the eastward ISM penetration depth was higher during periods of extended northern hemisphere warmth (e.g. the MWP) than during cool periods and vice versa.

The boundaries of LIA and MWP are defined in agreement with the timings given by Jones et al. [74] and within the periods of relative cold (warmth) in the East Asian temperature reconstruction by Osborn and Briffa [117].

On short (annual to multi-decadal) timescales, few studies systematically investigated the interactions between both sub-systems [129, 26, 187, 173] and all but [129, 79] focused on observational, reanalysis or model data. Nevertheless, the understanding of any system is fundamental to comprehending its links to other systems, and thus the the aim of this Chapter is to investigate the extent of interaction between the traditional ISM domain over continental India and the EASM domain over China. To this end I employ paleoclimate networks, based on significant association between proxy records of past climate variability. Palaeoclimate records come with particularities, when compared to data used in climate network studies up to now. They are heterogeneously sampled in time (1) and space (2) which, if ignored, leads to biased and possibly incorrect results. Previous climate network studies have focused on the analysis of gridded datasets, from reanalysis data [58, 150, 40, 42] or recent observations [54, 91, 92] and were thus restricted to the recent, observational period. Palaeoclimate records are, in contrast, spatio-temporally inhomogeneously distributed. However, due to the increasing number of (Asian monsoon) records published in the last decades [177], the spatio-temporal reconstruction of past climates might now become feasible [34, 177]. In difference to earlier climate networks, paleoclimate networks cannot make use of *direct* information about climatic parameters (e.g. temperature) and rely instead on proxy data that are usually irregularly sampled in time and space. Generally, fewer datasets are available the further back in time the analysis is extended. Also, much less paleoclimate data is available from India, compared to China. One option would be to include only datasets that span all time periods of interest and an equal number from both regions of interest (ISM and EASM domain). However, this would decrease the robustness and significance of the results. Therefore, it is necessary to strive to sample all regions consistently in order to retain comparability for different time slices, and include all records in the database where they meet the temporal sampling requirements. Possible bias effects should nevertheless be kept in mind for the subsequent analysis and need to be discussed.

5.1.2 Reconstruction of information flow – rather than physical state

Paleoclimate records of sub-decadal resolution are sparse in the ASM domain, but especially over the ISM region [177]. To include a maximal number of paleoclimate archives and improve spatial resolution and robustness of the estimates with increasing node numbers, the reconstruction of direct physical flows (which would limit us to using only precipitation or temperature reconstructions) is forsaken. Instead, I combine raw proxy records, not distinguishing between those that reflect precipitation or temperature, or both, as there is in many cases, there is a coexistence of factors that influence proxy variability [185, 20]. I argue that temperature and precipitation amounts over land covary, as the absolute water vapor content of air, and thus atmospheric flows, increases with their temperature [32, 163, 85] – provided liquid water supply is not limited. It is not sensible to claim that the general relationship between precipitation and temperature, especially in monsoonal and tropical climate, co-varies in a strict linear correlation sense either positively or negatively, but a (possibly nonlinear) association between the climate variables probably exists. Trenberth [163] found a negative correlation between monthly mean anomalies of boreal summer (MJJAS) surface air temperature and precipitation amount of reanalysis data (1979–2002) over much of India and China and state that “neither precipitation nor temperature should be interpreted without considering the strong co-variability that exists” [163]. Therefore, until a higher density of records for individual climate parameters is established,

I believe it is justified to use both to reconstruct the flow of dynamical *information*, measured by the extent of linkages, significant associations, between the time series of individual nodes. Combining different archives increases the robustness of the analysis against individual archive-specific biases, e.g., trees might provide information where stalagmites cannot or vice versa. In contrast to other analysis methods, every node retains its individuality in the network and its role in the final result, the network, can be assessed both visually (e.g. in force-weighted network representations) or quantitatively (by computing network statistics). Furthermore, should incompatibility be suspected, node removal is straightforward and does not require re-computation of the whole network.

5.2 Methods

Here, I briefly recapitulate the methods involved in the Paleoclimate network approach that were developed in the preceding Chapters.

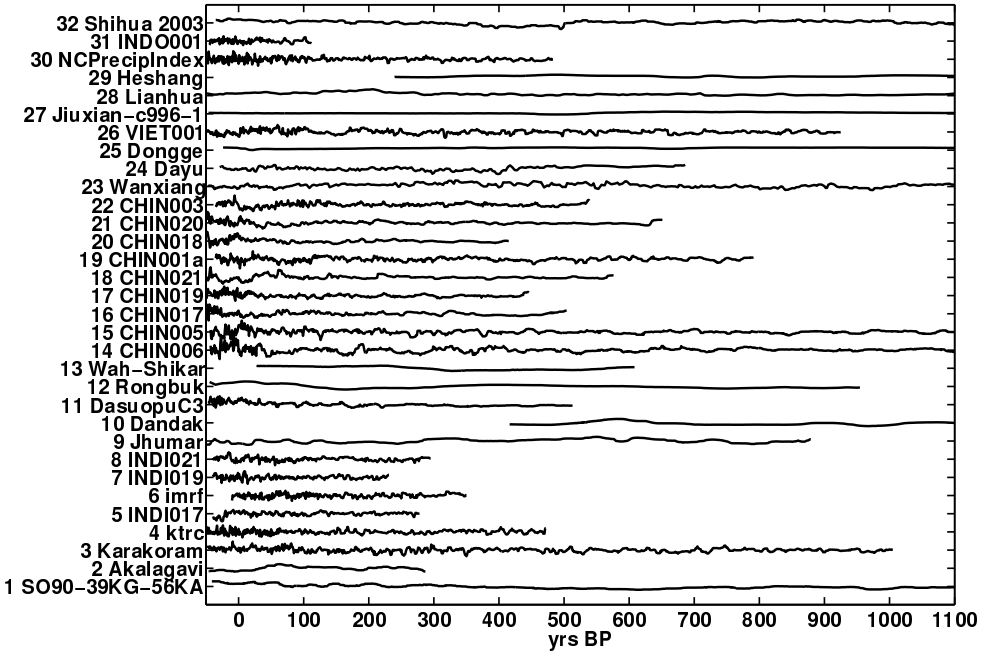


Figure 5.2: Temporal coverage of the Asian Monsoon records considered in the paleoclimate network (ensemble means over time, cf. Ch. 3) . While many records cover the recent past, few and fewer extend towards the end of the last millennium.

Time series representing paleoclimate conditions can be reconstructed from paleoclimate archives, e.g. objects, or entities, that accumulated information about past climate conditions. Typically, such proxy information is stored in structural properties of the archive, e.g. isotope ratios, grain sizes, or the amount of growth per year. Reconstructions of climate may focus on the absolute values of these proxies (reasoning for example that low values reflects cool temperatures), on time-dependent changes in properties of the proxy distributions. The truthfulness, i.e. much of the proxy variability is in fact due to climate variability, is determined via 'proxy calibration'. Ideally, such a calibration is established by comparing present-day local weather data to recent segments of the respective paleo record. It needs to be emphasized that processes of the very climate system that is sought to be reconstructed by use of paleoclimate records produces weather phenomena that ultimately lead to the recording of climate by the proxy: Stalagmites grow with

the supply of drip water, itself supplied to the cave via precipitation. Ice cores grow by adding firn to the top of the glacier, the necessary water transported to site via atmospheric circulation processes. If the growth conditions of the archive are unfavorable (drought, or re-routed atmospheric moisture supply) the observation of climate by the archive stops. Only resupply of 'growth material' could re-start the sampling process. Obviously living organisms, a single living tree, for example, will not be resurrected. but even in this case the record can be continued if young trees appear on site. Nevertheless, this climate-dependency of the archive growth inevitably implies that

- regular observation times can not be obeyed (irregular time series)
- the time of the proxy sampling climate is to some extent uncertain (chronological uncertainties)
- the location of the archive in space is dependent on local topology, geology, and climate (spatial sparsity and heterogeneity).
- and the climate processes sampled by the proxy may be representative for climate on a local and/or on a global scale. Or not at all. Consequently, proxy time series from the same, and different, regions need to be compared to discern inter-regional from local climate effects in a *replication test*.

All these above tasks and issues can be addressed within the framework of paleoclimate networks.

5.2.1 Quantifying dependencies for irregularly sampled time series

Similarity between time series in the geosciences is often inferred visually, using maxima, minima and general trends of a proxy time series (see for example [190, 76, 79, 127, 50]). Interpretation by eye however neither quantifies the similarity of patterns, nor is it foolproof[2]. Similarity measures, as presented in Chapter 2, evaluate statistical properties of two time series, thus quantifying the strength of their statistical association. While many algorithms require co-eval time axes of the time series, the methods developed in Chapter 2 can handle uneven spacing and avoid interpolation bias. The significance of such a similarity estimate can be tested by comparing the obtained value for the similarity to values observed for time series surrogates. The AR(1) surrogates employed here mimic the temporal sampling and the autocorrelation of the individual time series, but are mutually independent and uncorrelated (see Eq. 2.22 in Ch. 2 and [128] for more details).

[2] As a matter of fact, neither are statistics, but this shall be ignored momentarily for the sake of the argument.

Table 5.1: Records included in the paleoclimate network analysis. Records are listed from West to East. *LC* indicates **L**ayer **C**ounted archives, the asterisk * indicates that the respective age uncertainties were prescribed in this study.

No	Record	Lat. [°N]	Lon. [°E]	Archive	Proxy	Age uncertainty	Reference
1	SO90-39KG-56KA	24.8	65.9	marine	varve thickness	LC, abs. error, 10 yrs	[170]
2	Akalagavi	15.0	74.3	stal	δ^{18}O	LC, abs. error, 6 yrs	[182]
3	Karakoram	36.3	74.7	tree	rainfall	LC, % error, 1 yr*	[164]
4	ktrc	10.0	76.5	tree	ring width chronology	LC, % error, 1 yr*	[17]
5	INDI017	31.1	77.2	tree	rwl-crn	LC, % error, 1 yr*	[16]
6	imrf	12.5	77.2	tree	rainfall	LC, % error, 1 yr*	[118]
7	INDI019	30.4	78.2	tree	rwl-crn	LC, % error, 1 yr*	[16]
8	INDI021	29.8	79.2	tree	rwl-crn	LC, % error, 1 yr*	[76]
9	Jhumar	18.5	81.5	stal	δ^{18}O	LC, % error, 1 yr*	[146]
10	Dandak	19.0	82.0	stal	δ^{18}O	COPRA	[13]
11	DasuopuC3	28.2	85.4	ice core	δ^{18}O	LC, % error, 5 yrs	[161]
12	Rongbuk	28.0	87.0	ice core	Ca^{++}	LC, % error, 1 yr	[161]
13	Wah-Shikar	25.1	91.5	stal	δ^{18}O	COPRA	[146]
14	CHIN006	36.0	98.0	tree	rwl	LC, % error, 2 yrs*	[141]
15	CHIN005	37.0	98.5	tree	rwl	LC, % error, 2 yrs*	[141]
16	CHIN017	28.5	99.5	tree	rwl	LC, % error, 2 yrs*	[34]
17	CHIN019	29.1	99.6	tree	rwl	LC, % error, 2 yrs*	[34]
18	CHIN021	28.6	99.6	tree	rwl	LC, % error, 2 yrs*	[34]
19	CHIN001a	37.0	100.0	tree	rwl-crn	LC, % error, 1 yr*	[193]
20	CHIN018	29.2	100.0	tree	rwl	LC, % error, 2 yrs*	[34]
21	CHIN020	30.1	100.2	tree	rwl	LC, % error, 2 yrs*	[34]
22	CHIN003	38.2	100.3	tree	rwl-crn	LC, % error, 1 yr*	[194]
23	Wanxiang	33.0	105.0	stal	δ^{18}O	COPRA	[190]
24	Dayu	33.0	106.0	stal	δ^{18}O	COPRA	DY-1, [157]
25	Dongge	25.0	108.0	stal	δ^{18}O	COPRA	[175]
26	VIET001	12.0	108.5	tree	rwl-crn	LC, % error, 1 yr*	[24]
27	Jiuxian-c996-1	33.0	109.0	stal	δ^{18}O	COPRA	[25]
28	Lianhua	29.3	109.3	stal	δ^{18}O	COPRA	[35]
29	Heshang	30.0	110.0	stal	δ^{18}O	COPRA	HS4,[70]
30	NCPrecipIndex	37.0	111.5	historic + tree	JJA precipitation	LC, % error, 1 yr	[186]
31	INDO001	-7.2	111.8	tree	rwl-crn	LC, % error, 1 yr*	[44]
32	Shihua 2003	39.5	115.5	stal	temp	LC, abs. error, 5 yrs	[158]

5.2.2 Addressing chronological uncertainties

The correspondence between the position, layer or depth, in the archive and the time at which the proxy at this depth sampled climate processes is given by *chronologies*. Within the paleoclimate network approach, the uncertainties in the chronologies are forwarded into the analytical process by repeating it for different observation time realizations. From the ensemble of reconstructed observation times and the proxy measurements, an ensemble of the desired statistics is created (for more information, please refer to Chapter 3).

According to the treatment of the uncertainties within the presented approach I want to distinguish proxy records for which *layer-counting* was performed on one hand, from those which result from *age-modeling*, on the other hand.

Archives for which annual growth can be discerned are often *dated* by counting annual layered *varves*, or *rings*. Potential errors in these time-layer relationships can occur either because either a growth layer is missing in the archive, effectively contracting the apparent span of the record, or because a layer is double-counted. Usually the original investigators who publish a paleoclimate record indicate an absolute or relative counting error (e.g. [170, 182, 161, 158]). For all layer counted records included in the dataset for this Chapter I created 100 time scale surrogates, spanning from potential start-times to potential end-times. I allow for missing, but not double-counted years, assuming that each varve is always the result of one growth event and that they can be distinguished well from one another. This assumption is motivated by the fact that most of the records in the database are tree-based and that one tree will have only one ring per year, but not multiple [71]. For sedimented archives this might be a more – or less – valid assumption, as their growth is more irregular [170].

For **tree-ring records**, building a chronology involves so-called cross-dating of several, up to dozens, of trees. The probability that a year might go missing in the layer counting process for all trees in one site is low, if one assumes that this occurs randomly in time, at a rate of about one percent (personal communication, G.Helle). If, however, trees stall growth for climatic reasons [as described, for example in ref. 98], all trees at a given site could be affected. Additional – and large – uncertainty might arise if the start and/or the end of active tree growth have to be inferred from radiocarbon dating[3]. In this present study I assumed a relative chronological error of one percent for readily published tree chronologies (*crn*), and two percent for chronologies that I built from published raw ring width data (c.f. Rehfeld et al. [129] for information on how these *rwl* chronologies were built). Tree ring width chronologies (indicated by *rwl-crn* in table 5.1) were assigned a 1% counting error to obtain observation time ensembles. Raw tree ring width series (*rwl*) were assembled into chronologies by first detrending the individual tree series with a 50-year Gaussian kernel smoother to remove youth bias. Then the individual ring width series were standardized and averaged to obtain a site-specific chronology for the corresponding years.

Proxy record ensembles from age modeling rely on the accumulation, or growth, time estimates from dating techniques. Depth-age relationships for non-laminated archives, such as many speleothems, are, for example, often established via radiometric U-/Th dating. For these records in the dataset the published dating tables and proxy measurement depths were combined using the **COPRA** algorithm (c.f. [21] and Chapter 3). The COPRA-modeled depth-age relationships can be found in Appendix 6.

[3] personal communication, G. Helle

5.2.3 Test spatio-temporal dependencies using paleoclimate networks

At the core of the paleoclimate network approach (c.f. Chapter 4 for more details) lies the question, whether or not two proxy time series are significantly similar. The time series in the dataset are understood as *nodes* that are spatially embedded and located in the domain of interest. If the climate time series between two nodes are significantly similar, these nodes are considered to be *linked*. This test for significant statistical association is repeated for all pairwise node combinations, similarity measures and time series ensemble members, from which a mutual *link strength* between each tuple of nodes is obtained, factoring in time uncertainty, time scale irregularity and sparsity.

5.2.4 KIMONO: A toy model of spatio-temporal dynamics in the ASM domain

Considering all the uncertainties involved in paleoclimate reconstructions, it seems justified to ask: Can dynamical information be gleaned from such a sparse and heterogeneously sampled dataset? To test this – and a possible connection of the Indian Summer Monsoon influence on the Chinese monsoonal domain – I employ the KIMONO toy model for ASM dynamics that was presented in Chapter 4 is employed. Hypothesizing that higher regional temperatures result in a a relatively stronger ISM than EASM circulation, the proxy measurement values at each location in the original dataset are replaced by the KIMONO output at these points. Please note that this way the spatial and temporal sampling of the dataset is preserved, only chronological uncertainties are not.

5.3 Data

The data in the considered dataset include proxy data from between 66-116°E and −7 to 39°N (Fig. 5.1). Exact position and details on the records, the corresponding proxies and the age uncertainty considered are given in table 5.1. To be included in the dataset, the proxy record had to be published and publicly available. The dating tables and the proxy measurement depths were partly obtained from the publications, partly supplied by the original investigators. Furthermore the proxy record had to cover at least one of the 12 overlapping 300-year periods with at least 50 observations. The number of datasets that fulfill these requirements decreases over time (Fig. 5.3), and generally fewer datasets are available from West of 100°E compared to East of this imaginary border. For simplicity, in the following all locations positioned West of 100°E are considered to lie in the domain of 'India', all those further East are considered to be located in 'China'.

5.4 Results

I performed a sliding window paleoclimate network analysis with a window width of 300 years and an individual overlap of 75%. This resulted in 12 individual time slices for a start year of -50 yrs BP(equivalent to AD 2000) up to 1100 yrs BP. For each data location a time series ensemble with 100 members was considered, for which networks were constructed individually via similarity estimation and significance testing. Note that while for the actual paleoclimate data the time series ensemble consists of one set of proxy observations and 100 possible time axes (denoted *LHens* in the following), the ensemble for the temperature-forced KIMONO model consisted of one (the original) time axis with 100 realizations of the KIMONO model output (*KinoTemp*). Additionally, the network was reconstructed for the *original*, published and time-uncertain chronologies (*LHunc*). For the zero lag similarity estimates a 90% significance threshold was chosen and individually tested with 100 AR(1) surrogates. The average link density, i.e. the

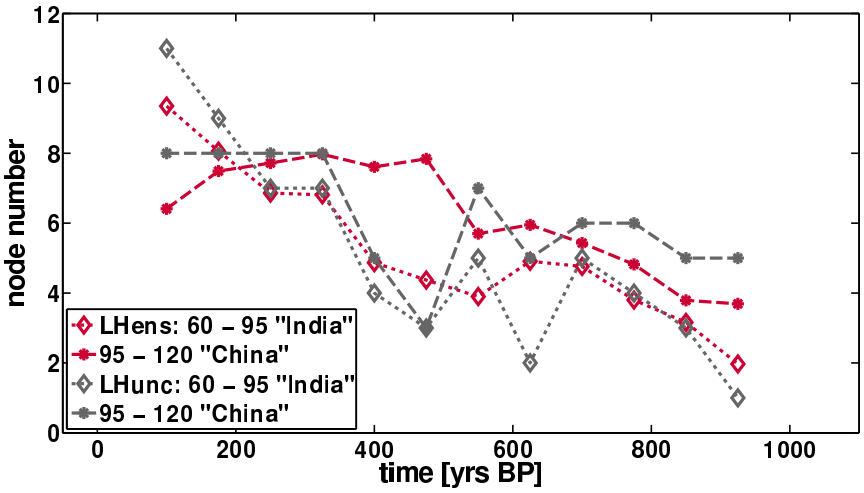


Figure 5.3: The number of datasets per time slice which fulfill the minimal requirements (50 observations in a 300-year sliding window), decreases as the reconstruction goes further back in time. Numbers are graphed with respect to the center point of the time interval. The varying time axis within the *LHens* run, in red, prevent the sharp decrease in includable datasets for the age uncertain *LHunc* run, shown in grey.

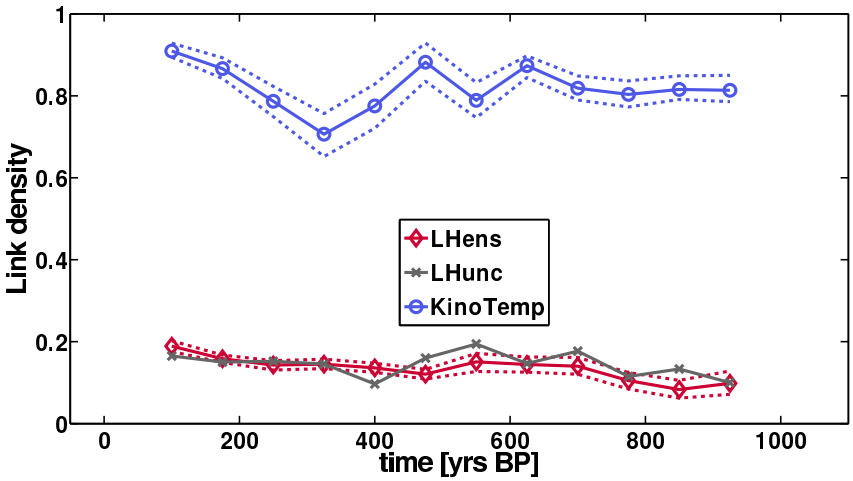
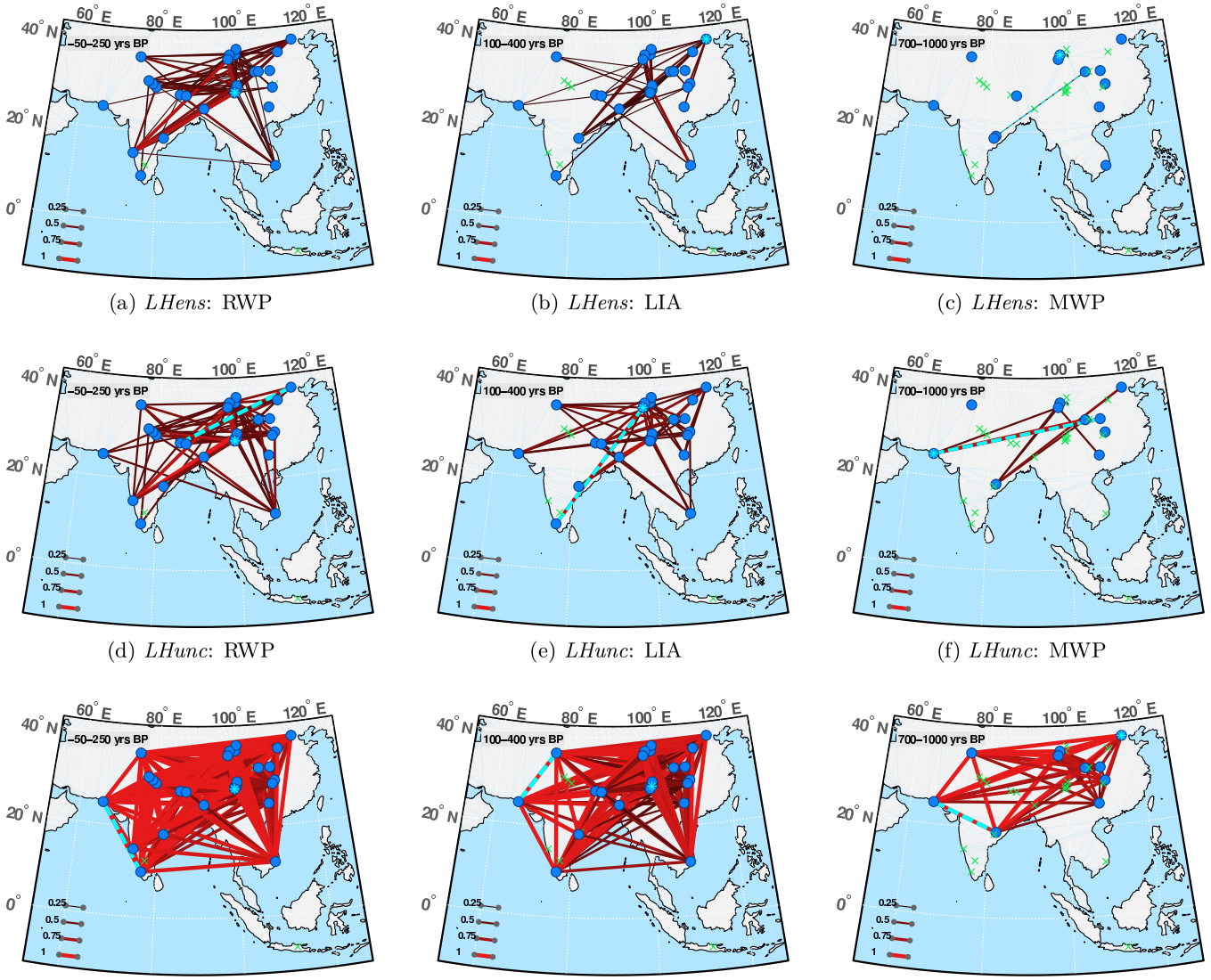


Figure 5.4: The average link density is not affected significantly by the decrease in node numbers. Please note that due to the age uncertainties addressed in the ensemble runs the number of nodes per time slice and area, is given as an ensemble average – as is the corresponding link density.

averaged fraction of *realized* links in one time slice, observed in the networks is depicted in Fig. 5.4. For the *LHens* ensemble it remains constant at around 0.2 from 100 yrs BP to 700 yrs BP, when it drops to 0.1 at around 950 yrs BP, parallel to the decrease in network size, as seen in Fig. 5.3. Considering the 90% significance threshold, a link density of 0.1 is also what would be expected for mutually uncorrelated time series.

5.4.1 Reconstructed paleoclimate networks for the recent past, the LIA and the MWP

Figure 5.5 shows snapshots of the network evolution for the time periods of the MWP, the LIA and the RWP for the data ensemble run *LHens*, the original data *LHunc* and the model run *KinoTemp*. For all three runs the same significance threshold of 90% confidence was chosen. The link strength for *LHunc* are determined by the fraction of similarity measures ($gXCF$, $iXCF$, gMI, iMI and ES) for which significant similarity was observed. By contrast, the link strength for the ensemble runs, *LHens* and *KinoTemp* were determined from both similarity measures and ensemble realizations, thus broadening the statistical base.



(a) *LHens*: RWP (b) *LHens*: LIA (c) *LHens*: MWP

(d) *LHunc*: RWP (e) *LHunc*: LIA (f) *LHunc*: MWP

(g) *KinoTemp*: RWP (h) *KinoTemp*: LIA (i) *KinoTemp*: MWP

Figure 5.5: Paleoclimate networks for the *LHens* data ensemble run, the *LHunc* reconstruction on the original chronologies and the *KinoTemp* model run, for which the temporal sampling of the *LHunc* dataset was used. Only links with a link strength exceeding the 0.1 significance level are plotted. Nodes which do not provide sufficiently high resolution time series.

Comparing the observed networks with respect to the different time slices it is apparent that the overall link strength in the MWP is lowest, and that in the RWP is highest for all data types. Furthermore, the observed links in the model run *KinoTemp* are strongest, those in the data ensemble seem to be weakest. In Figs. 5.5 (a)–(i) the strongest nodes, node strength equaling the sum of link weights connected to the respective node, are indicated by light blue stars, the strongest links by light blue dashed lines.
In the following I want to take a closer look at the reconstructed networks for the three time periods, MWP, LIA and RWP. In the **RWP** the strongest nodes (CHIN021 for *LHens* and *LHunc* and the nearby CHIN017 for *KinoTemp*) are consistently at the center of the spatial domain. While for the model run the links are consistently significant, for the *LHens* and *LHunc* runs a triangle, formed by strong ($p > 0.8$) links stand out, connecting the South Indian Akalagavi record with North Chinese tree ring sites (CHIN005,CHIN019,CHIN020).
For the **LIA** networks the influence of age uncertainty becomes more apparent: The strongest nodes are still found in the Chinese network domain (Shihua, CHIN006 and CHIN018) but compared to the *LHunc* network the *LHens* reconstruction seems thinned out. The strongest link for LHunc, connecting CHIN006 and the Kerala tree ring chronology ktrc is not preserved if age uncertainty is introduced.
LHunc In the **MWP** overall only one link exceeding the significance threshold is robust under consideration of age uncertainties. The strong link between Wanxiang and SO90-39-KH-56KA in the *LHunc* network disappears completely. Surprisingly, the North-Chinese record from Shihua cave is the strongest node in the *KinoTemp* network, although the prescribed ISM temperature forcing (c.f. Fig. 5.6) should dominate the Western part of the network.

5.4.2 Network measures for the last millennium

Regional degree ratio

Figure 5.6 gives the estimated network measures for the 12 sliding windows. Splitting the network into nodes in China vs. nodes in India the degree ratio indicates how homogeneous the distribution of nodes across the network is. The model trajectory remains consistently above the gray dashed line indicating homogeneity and varies little over time. This showing that the average degree in India is consistently *higher* than that in China. In contrast to this, the results for *LHens* and *LHunc* networks lie below this line for the most part, indicating higher node-strength over the Chinese subdomain. For the most part the expected trajectory for the *LHens* agrees with the *LHunc* estimate. At the onset of the LIA, however, for the time slices around 400 and 475 yrs BP, the degree ratio drops dramatically for LHunc. An inspection of the reconstructed networks for these times (c.f. Fig. 14 in Appendix 6) clearly shows the reason for this: Although nodes in the Indian domain exist in these time slices, no significant links amongst them exist! By contrast, during this time mutual links are abundant in the Chinese domain. Another drop in the degree ratio is observed at the end of the LIA, and it is even more pronounced then for the *LHens* as for the LHunc runs, where similarly many mutual links in the Chinese subdomain are observed, and few in the Indian subdomain. Also, at this point the dashed lines, indicating 65% the values for the distribution of the degree ratios of the ensemble does not encompass the homogeneity line.

Cross-link ratio

Fig. 5.6 gives the relation of cross-link probability, domain connectivity, to the overall link strength. Here again, the line at unit cross-link ratio indicates homogeneity. For all time slices, this line is encompassed by the confidence intervals for the *LHens* ensemble, indistinguishable from random networks. The estimate for *LHunc* lie within the *LHens* confidence bounds for the

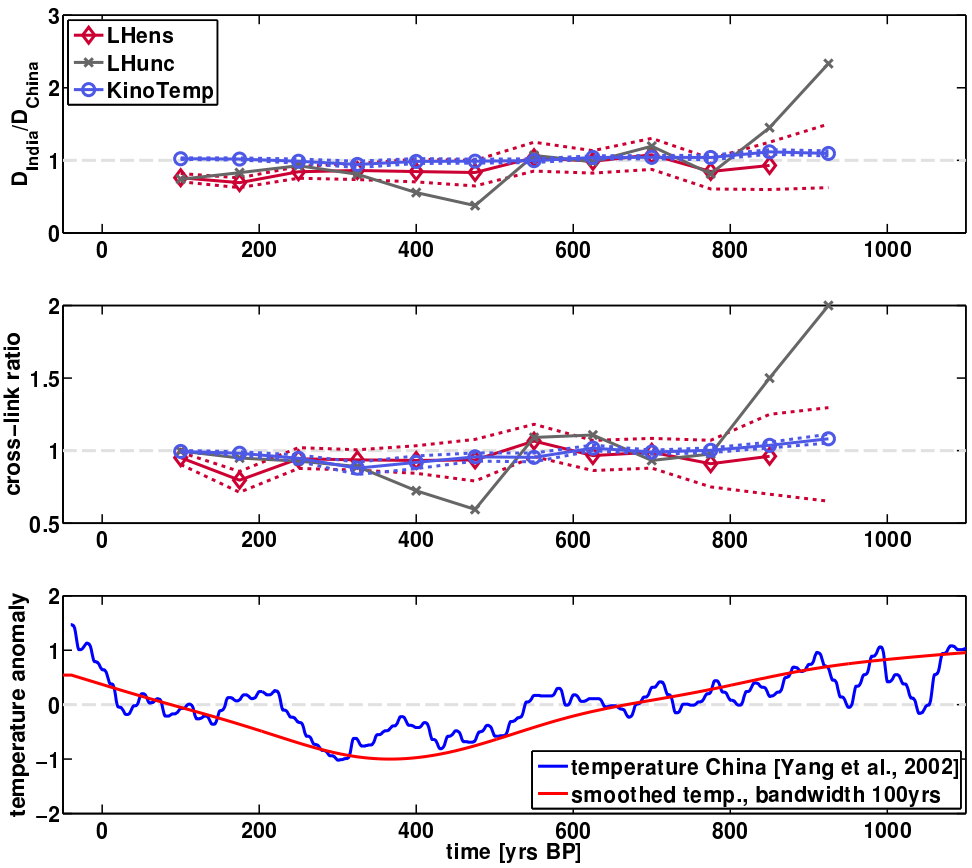
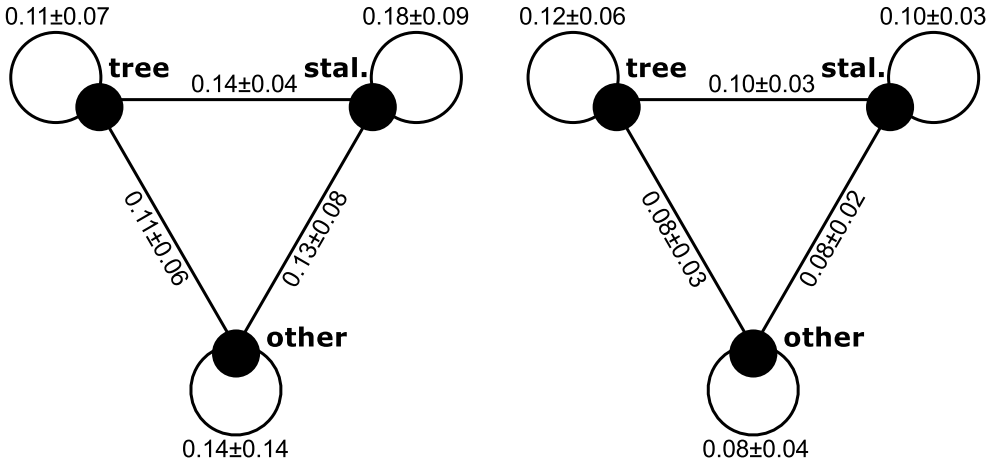


Figure 5.6: **Top:** Regional node strength. **Middle:** cross link ratio. **Bottom:** Reconstructed South Asian temperature [184]. The gray dashed lines indicate the equilibrium lines for equal degree in the subnetworks, equal cross-link to overall link strength resp. indicates the mean temperature.

most part, except at 400 yrs BP where it lies above. This outlying value ties with the previously noted sharp drop in the Regional Degree ratio, and results from the absence of links within the ISM domain. This feature is, however, not picked up by the *LHens* ensemble mean. The model trajectory starts at an above-unity value at 925 yrs BP, then remains constant until it drops between 600 and 400 yrs BP, rising to homogeneity towards the present.

5.5 Discussion

5.5.1 Potential archive-dependent or local bias effects



(a) Link strength: original chronologies

(b) Link strength: Age-certain ensemble

Figure 5.7: Archive-dependent bias effects due to different climate sensitivities could distort the paleoclimate network. Three classes of archives considered here, trees, speleothems and other archives (marine sediment, ice core, historic documents, etc.). The mean and standard deviation of the frequency of occurrence of links amongst these archive types are given along the edges of the graph. For Fig. (a) the original chronologies were considered, and age uncertainties ignored. Including the age uncertainties for Fig. (b) decreases the observable link strength to the significance level.

If the archives recorded climate in a systematically different way, their auto-link strength should be higher than the link strength to other archives. To test this, I estimated the average link strength and its standard deviation for nodes of the classes *tree*, *stalagmite* and *other* (e.g. ice core, sediment core, historical documents), based upon the link strength matrices for all time windows of *LHens* and *LHunc*. The results are given in Fig. 5.7. The stalagmite records included in the analysis show the highest link strength amongst each other (0.18). All auto-and cross-archive link strength are above the link strength 0.1 which would be expected for arbitrary, random networks. However, for all of them the one σ uncertainties still *embrace* the link strength 0.1. This means that while there is reason to believe that there could be a common linking factor (climate!) for some links, it does not mean that this is generally true. The link strength for most archive types drops below the critical value of 0.1 if age uncertainties are considered (Fig. 5.7b), although the auto-link strength for trees increases slightly. This could indicate that the considered uncertainties actually do make the cross-comparison of tree chronologies more efficient.

A clear decrease of link strength with link distance is observed in climate networks constructed

from daily temperature and pressure fluctuations [11, 165], indicating localized climatic effects. Similarly, also in spatially embedded transportation and biological networks [8] the link distance is an important aspect – as "wiring costs" increase with link length. Such effects in climate networks indicate the importance of convection and diffusion processes versus long range teleconnections on short, intra-annual timescales. In contrast to this, the strength for two records to be linked in the $LHens$ paleoclimate network does not depend on their mutual distance (c.f. Fig. 11 in Appendix 6). This could indicate that local climate variability is less represented in the paleoclimate archives than global factors, which in turn also supports common driver hypotheses instead of local convection-based dynamics for ASM dynamics. The asynchronous evolution of the Asian monsoon systems on centennial scale, as postulated by Wang et al. [177], can therefore not be supported on the decadal scale. Such a discrepancy could be resolved using observational or model data. Unfortunately, for meteorological data, decadal-scale variability is difficult to assess due to the short period for which they are available. Such effects could potentially be investigated using millennial GCM simulations [172].

5.5.2 Chronological uncertainties as a limiting factor

The information that can be gleaned from the paleoclimate network configuration as such is limited to dynamics on certain time scales by several factors:

1. The window width W that dictates how well potential time-dependent dynamics of the Earth system can be resolved. For example, the LIA had a duration of $\approx$ 300 years, therefore the analysis window in a sliding window analysis should not be larger.

2. The number of time points in the window W, N_{obs}, should be at least 50. This dictates the average time resolution of the paleoclimate network, Δt. From the estimator side, as many observations as possible are needed. On the data side, the more strictly this requirement is enforced, the fewer datasets can be included, and this reduces the spatial resolution.

3. The persistence time of the time series of the order of 7-10 years [128] induces serial dependency and in principle increases the required length of the time series for the estimators [109].

4. The age uncertainty which increases the further the analysis is taken back in time (Fig. 5.8)

5. Detrending of the time series using a nonlinear Gaussian smoother of bandwidth W_b results in a high-pass filter.

The actual time range in which therefore potential variability can be interpreted is limited by these high and low-pass filters. For this study it is in the order of a decade to half a century.

Contemplating Fig. 5.7 and Fig. 5.8, it becomes obvious where one of the main challenges for successful paleoclimate reconstruction lies: In the age uncertainties of the records. While speleothem records show a higher auto-link strength if their original chronologies are used, this drops to complete insignificance when age uncertainties are considered. This indicates that a) the age uncertainties for these records are large enough to destroy a mutual similarity and b) it could be that the chronologies of non-laminated records, as published, are tuned to be similar to other stalagmite chronologies. This could, of course, also affect other such archives, e.g. ice cores, which are not abundant in the present dataset. The observed drop in link strength is also visible in the network topologies that were shown in Fig. 5.5: For the MWP only one weak link is observable when age uncertainties are considered, compared to several if this is not the case. Paleoclimate records that span into the MWP (c.p. Fig. 5.2 and Table 5.1) tend to have lower resolution and higher uncertainties than in the more recent periods. Considering age uncertainties leads to a

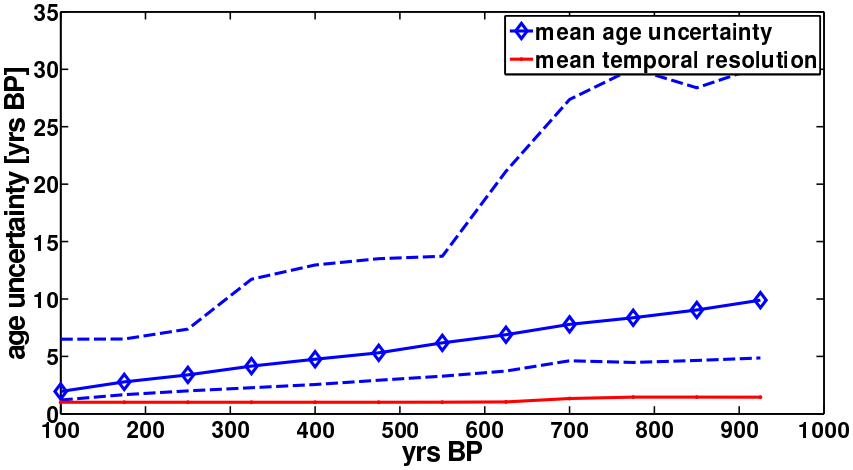


Figure 5.8: The mean age uncertainty (blue line) in the dataset increases over time. Dashed lines give the .25 and .75 quantiles. The mean temporal resolution in the datasets is given in red.

blurring of characteristic phenomena observed in the network measures. Which of the signals – the smoothed ensemble version, or that of the age-uncertain single realization – is in fact closer to the truth can not be determined at the moment. In this analysis time slices of 300 years, with time series of sub-decadal resolution were considered. Assuming a 1% relative age uncertainty due to counting errors, the age error is in the order of magnitude of the sampling minimum if the analysis is taken back even only few centuries. This increase in uncertainty could in principle be compensated by a larger number of records considered for the analysis. For the ASM domain, at present, this is, unfortunately not feasible and precludes reconstruction of decadal-scale dynamics for the last millennium.

5.5.3 Influence of temporal changes on the spatial node distribution

The relevance of the *KinoTemp* experiment ties, in part, to the exclusion of strong bias effects due to the spatial distribution of nodes. By resampling the model results for each paleoclimate record location on the respective time axis of the record it is possible to assess potential bias effects of irregular sampling in both temporal and spatial domain. To negate such effects the changes in the forcing of the model (Fig. 5.6) should be recognizable in the observed network measures (c.f. Chapter 4 for the effect of the forcing on the network measures). However, the cross-link ratio seems to be unaffected by the forcing changes (Fig. 5.6). This could be due to the extremely high link density, as the spatio-temporal synchronousness implied in the model persists through the temporal resolution changes. Future model runs could aim at lower, more realistic, proxy sensitivity, towards 30-50 instead of 90%, as for example in the GCM-based pseudo-proxy experiments by von Storch et al. [171]. The amplitude of the cross-link ratio change for Fig. 5.6 that ties with the changing influence of the longitudinal ISM component is small compared to the changes found in the real datasets. While this might be due to the inadequacy of the model to represent actual ISM dynamics, or the inappropriateness of the selected network measures, it also indicates that, with regard to model dynamics, the spatial and temporal sampling is sufficient for these periods to reflect changes in spatio-temporal dynamics. To test the suitability of the sampling for selected time periods, the model should be run under different forcing for each period. If the observed networks are not well distinguishable by network measures the spatio-temporal sampling is inadequate. Similarly, the currently not considered age uncertainty could be introduced in the model run.

5.5.4 Temperature as a driver for Indian Summer Monsoon influence on climate in China?

KIMONO model output, was generated for the aforementioned spatio-temporal configuration. A reconstruction of South Asian temperature was used as forcing parameter for the model [184]. The dominant feature in the temperature reconstruction by Yang [184] is a slow dominating drop to pronounced cold temperatures in the LIA around 400 yrs BP and a quicker transition out of this minimum towards the present day. A 100-year smoothing of this reconstruction was used as linear forcing input (c.f. Sect.4.2and Fig. 5.6). Comparing the observed trajectories to those estimated for the data, no direct dependency of cross-link ratio or degree ratio on the changing forcing parameter is observable. While this could be due to uncertainties in the temperature reconstruction (which were, unfortunately, not given!), it is also not likely that local factors could dominate ASM strength alone. The global influence of factors such as NH temperature, ocean circulation or glaciation could acting as 'common drivers', synchronizing the local climates across the ASM domain. Such phenomena are not simulated within KIMONO, as it is a model for information transfer due to physical mass transfer. Still, it has been shown that ASM strength and local temperature follows NH insolation more closely than mean NH temperature [190]. However, the one sharp and significant drop in the estimated cross-degree falls into a time where the transition out of the cold LIA towards the warm present day reverses briefly. The sharp drop observed for *LHunc* around 650 yrs BP might relate to the mid-14th century monsoon weakening that 'is not likely to be the result of [a] solar shift' [190], and although solar forcing certainly plays a role in monsoon dynamics, other factors contribute. Considering the age uncertainties in the dataset it is not possible to discern a clear picture of how to improve the discrepancy between the model dynamics and the paleoclimate dataset. Such improvement could target the assumed response equations to forcing (Eq. 4.18), missing components such as the Westerlies, or teleconnections not modeled due to the assumption of direct physical flow.

5.6 Summary

To conclude, I find that

- the distinction of local temperature-change induced spatio-temporal patterns in sub-centennial to decadal ASM dynamics is precluded by the considerable age uncertainties of the same order of magnitude. The observed paleoclimate network measures are, however, not consistent with a linearly temperature-modulated synchronization of the ASM domain via a direct physical, convection-based, longitudinal ISM flow as modeled in KIMONO. This could indicate a more relevant role of other atmospheric circulation systems (e.g. the Westerlies or the EASM, or that non-linear response functions need to be considered for the ASM model.

- Age uncertainty is the most relevant challenge that has to be met in order to enable systematic interpretation of the Paleo-ASM on sub-decadal timescales. Time-scale irregularities are not a hindrance, and spatial heterogeneous placement of data sources inhibits sensible interpretation of network measures if less than four records per sub-domain are provided.

- General archive dependent bias effects are not observable, thus supporting the reconstruction of information, instead of physical parameter flow, as the dynamical information they convey is not, by default, biased. No relationship between link distance and link strength could be observed, and strong links are found both at close proximity as well as thousands of kilometers apart. This indicates that decadal-scale dynamics of the ASM are most prob-

ably not internally, but externally generated and ties with the global aspect of the monsoon circulation.

6 Discussion and Outlook

Discussion and Outlook

The work presented developed methods for the reconstruction of paleoclimate dynamics. Inspired by the increasing popularity of complex network theory and application, I put forward the paleoclimate network approach as a means to investigate spatio-temporal dynamics and transitions. The central questions addressed in the course of this thesis focused on the key challenges of paleoclimate reconstruction: Considering age uncertainty and temporal inhomogeneity – how can similarity between paleoclimate proxy time series be quantified? What effect does the heterogeneous distribution of paleoclimate archives have on the estimation of network diagnostic measures?

In the following I summarize and discuss the findings of the previous Chapters.

Similarity estimators for irregularly sampled time series

Finding and testing efficient and robust similarity estimators for irregular time series was at the heart of Chapter 2. The principal idea behind the methods I adapted and put forward is to use the time series data *as they are* and to avoid interpolation wherever possible. The Gaussian-kernel-based cross correlation estimator shows clear advantages compared to the traditionally employed standard methods using interpolation. While cross correlation functions (CCFs) are estimable efficiently for short time series, mutual information (MI) estimators are at the extreme lower end of their data requirements due to inherent bias effects. Therefore, although the local Gaussian kernel-based MI reconstruction scheme improves the robustness against varying temporal sampling, more research should be devoted to the development of adequate estimators in the future. The trio of linear and nonlinear association measures is completed by an event synchronization function (ESF). Both CCF and MI center on the joint distribution of the data analyzed and, in principle, require coinciding, bivariate, observations. By contrast and as defined here, the ESF is computed using the relative position of extreme events, or excursions, in the data. Bivariate observations, as such, are not required for the method. The provided methods are adequate to quantify similarity amongst heterogeneously sampled time series and present a clear improvement over standard methods. The introduced concept of a link strength summarizes the statistical significance of estimates of statistical association flexibly.

Recently, a Gaussian-kernel-based Granger causality estimator was implemented and tested analogously to the here-presented cross correlation estimator [5]. Thus future improvements could be sought in the development and integration of methods that are robust with respect to sampling irregularity and can, additionally, infer the direction of a potential coupling mechanism. This includes, apart from the aforementioned Granger causality, for example approaches based on recurrence networks [48] and recurrence plots [100, 84], but could also apply to an adaptation of the event synchronization function. Multivariate extensions, for example based on distance measures [86] could be especially useful to compare naturally multivariate paleoclimate proxy datasets such as pollen data results from different models at different resolution.

Age uncertainty and similarity estimation for time series

Time is a variable that has to be reconstructed for paleoclimatic purposes. In Chapter 3 I combined a numerical approach [21] to modeling accumulation histories for dated archives with synthetic paleoclimate archives and benchmark tests for regular and irregular sampling (as in Ch. 2). Comparing short time series with up to 200 observations, I found that for CCFs and MI the largest contribution to uncertainty in their similarity estimate is the introduced age uncertainty. If the time series are interpolated for standard estimators, this changes and the irregularity of the sampling becomes the largest contributor. By contrast, the ESF can be estimated more efficiently, when the sampling is irregular than when it is regular, and it has the best overall performance. Although this is at first surprising, it arises from the increased amount of information on short timescales that is available if the sampling is irregular, and that is counted efficiently by this estimator. Additionally, the ESF by construction does not require the processes to be observed coevally.

Forthcoming work on dependency quantification for age uncertain time series should also include nonlinear benchmark models, both for the time series as well as the synthetic proxy data. The flexibility of the current benchmark test with regard to the specific estimator being tested, could make it especially useful to test for example the limitation of frequency-domain information (as gleaned from wavelet or power spectra) in the presence of age uncertainty, or the significance of trends [52, 112]. Finally, the suitability of the ESF for age uncertain data indicates that it could be applied to infer the impact extreme events not only in the instrumental era [91] but also in the more distant past. Embedded in the paleoclimate network approach, this could help to visualize and understand the spatio-temporal synchronousness of extremes in past climate, e.g. the impact of Dansgaard-Oeschger events [51, 30, 113].

Determining changes in spatio-temporal dynamics from unevenly distributed nodes

Pinpointing the spatial extent of past climate changes based on sparse and heterogeneous data is difficult. In Chapter 4 I used a simple semi-empirical model of convective flow extent to model dynamical information transfer in the Asian monsoon domain. As presented, the KIMONO model has simple dynamics that, using a single forcing factor, transitions from a state with two separate synchronized regions and dominant vertical flow to a state with one large region of influence and longitudinal flow. I used this model to assess how different complex network measures show the spatio-temporal changes in the dynamics. Also, the model domain can be sampled at different locations. The tests conducted were based on both a regular reconstruction grid, and the locations for which high-resolution paleoclimate data are available in the Asian summer monsoon domain. I found that the changes in the dynamics result in robust changes in the diagnostic network measures, but that nontrivial bias effects occur if the spatial sampling scheme is changed. On one hand, the spatial distribution of nodes might amplify certain signals if they are situated in a dynamically important region. On the other hand changes might have to be more pronounced if the sampling locations lie in less important regions. Caution should therefore be exerted before network measures are interpreted quantitatively.

Still, in many applications [34, 176, 177, 190, 70] heterogeneous paleoclimate information is used to infer the spatial extent of past climate phenomena. Spatio-temporal transitions are well-reflected in paleoclimate network measures and should be compared to spatio-temporal reconstruction techniques, such as EOFs and multivariate SSA. Such pseudo-proxy benchmark tests could also help in the identification of key regions from which paleoclimate records should be obtained, using the combination of a simple model of the investigated dynamics, and the detectability improvement of diagnostic measures.

Testing temperature dependency of the Indian summer monsoon strength

In Chapter 5 the aforementioned methods were combined to test for a potential temperature modulation of the inter-annual Indian summer monsoon extent, as proposed previously [129]. Small paleoclimate networks were constructed for sliding time windows over the past 1100 years. Few of the paleoclimate datasets span the whole time period at sub-decadal resolution, therefore the numbers and locations of nodes change continuously over time. Age uncertainty was included in the analysis by considering ensembles of time series for each node, and therefore ensembles of reconstructed networks were obtained, quantified and interpreted. In parallel, the KIMONO model was run to generate pseudo-proxy records for the same observational period and spatio-temporal sampling. Forcing the KIMONO model using local temperature reconstruction for China, I simulated a strong change in underlying spatial dynamics, from bisected lateral flow to a completely homogeneous longitudinal flow. This is reflected in the network measures, but the amplitude change of this transition is small due to the spatially and temporally heterogeneous sampling. Comparing the results for model and paleoclimate data, I found that a) sub-decadal resolution reconstruction is hampered by age uncertainties even if the analysis is not taken back far in time, and b) for the data, the observed network measures show inconclusive results regarding possible temperature-dependent changes in the relative importance of longitudinal vs. latitudinal ASM flow when age uncertainties are incorporated in the analysis. This could be due to the rather conservative estimates for errors in tree ring chronologies, as most dendrochronological studies assume there to be no errors. Also, KIMONO currently only considers local convection as a means to synchronize regions. The Asian monsoon system, however, is not separate from global dynamics and solar forcing. The observed network with homogeneous link strength could indicate, for example, that dynamical changes were occurring consistently all through out the domain and that common drivers on a more global scale are the most important cause for apparent time series similarity. Another reason for the inconclusiveness of the results could be the use of raw proxy data from different archives instead of reconstructions of specific climatic variables. However, if, e.g., ring width in a tree-ring chronology in India shows statistically significant similarities to a stalagmite record in China, they share some amount of dynamical information on climate that affects both. The use of heterogeneous archives did also not impact the analysis, because archive-dependent bias effects could not be observed. The predominant reason why the reconstruction of paleoclimate dynamics is difficult, are the sparsity of data and inadequate dating control, as already Mayewski et al. [102] cautioned. However, to improve the understanding of the situation, local and global sources could be integrated in the KIMONO model. The network results indicate that the time period up to 300 BP could yield more robust results. Beyond this, the age uncertainties and spatial sparsity prevent conclusive interpretation on sub-decadal timescales as the requirements for the estimators (temporal sampling density), age uncertainty, and the extent of interesting dynamical features (e.g. $\approx$300 years for the LIA) conflict. Such data availability issues could be less harsh if other regions of the world, such as Europe or North America, were considered, and where more paleoclimate data is available [73]. Until a comparable spatio-temporal resolution exists, analyses should be restricted to a wider temporal windows and lower temporal resolution. Still, using the concept of paleoclimate networks, teleconnections of the Asian monsoon system such as El-Niño Southern-Oscillation, or its ties to Europe mediated by the westerly continental winds could be investigated, considering them as separate, but interacting paleoclimate networks. Similarly, the influence of potential forcing mechanisms can be assessed by quantifying the link strength between the ASM domain and, for example, records of solar variability [152], volcanic forcing [192], or oceanic coupling via sea surface temperatures [27, 126, 162].

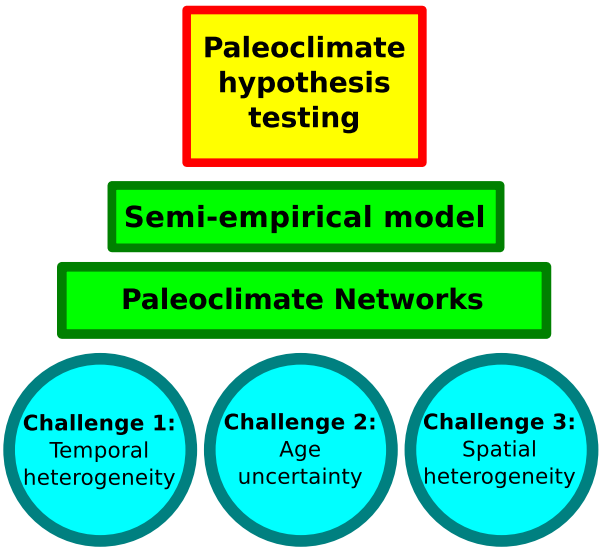


Figure 6.1: Within the paleoclimate network approach, the key challenges in paleoclimate reconstruction, age uncertainty as well as temporal and spatial inheterogeneity can be flexibly addressed. Together with (semi-empirical) modeling the approach can be used to test hypotheses on spatio-temporal paleoclimate dynamics.

Conclusion

The paleoclimate networks put forward in this work is a flexible framework to address the key issues of paleoclimate (Fig. 6.1): temporal heterogeneity can be handled efficiently by adapted similarity estimators. Age uncertainties can be propagated through the analysis by considering ensembles of time series. And in itself the approach does not require an underlying spatial grid, since the effects of spatially heterogeneous sampling can be characterized in combination with semi-empirical or fully dynamical models. The adapted similarity estimators, and the benchmark tests to infer their suitability, are of relevance and partly already in use in other disciplines in their own right [106, 5]. In order to confirm a potentially superior suitability of the paleoclimate network approach to infer spatio-temporal changes from sparse and heterogeneous paleoclimate data, its results should be compared to that of traditional methods such as spatiotemporal EOF analysis and multivariate SSA. Still, using it spatio-temporal transitions can be detected in heterogeneously sampled datasets as long as age uncertainty is not too large and the dataset not too sparse. Thus, accepting the heterogeneity of information the Earth provides about its own past might be as good as, or better than, forcing it onto regular grids in time or space.

Appendix A

1 Derivation of KIMONOs spatial variance distribution from the Advection-Diffusion equation

In this section we[1] develop the solution of the Advection-Diffusion equation for the simple toy model for Asian Summer Monsoon (ASM) atmospheric circulation extent. Observations and analysis of paleoclimate data have shown that the core region of the ASM is influenced by main branches of the Indian Summer Monsoon (ISM) and the East Asian Summer Monsoon (EASM). While the ISM influence is primarily a directed West-East flow, the EASM influence can be interpreted as a predominantly South-Eastern atmospheric circulation. Over the centuries the respective strengths of these influences seem to have varied significantly. We hypothesize that the varying strength of the longitudinal ISM source results in varying extent of atmospheric flows into the intermediate domain in which ISM and EASM compete. Testing a potential modulation of the ISM strength by external factors (i.e. global or Northern Hemispheric temperature) directly would require extremely developed coupled global circulation models (GCMs) with adequate and well understood physics. Ground truth to verify these model simulations would have to be found in paleoclimate proxy data. These in turn are riddled by uncertainties (age uncertainty, proxy representativity, temporal and spatial sparsity), and thus need to be verified themselves.

In this context we propose the use of our small statistical model in combination with a complex network approach to investigate the potential of the currently available paleoclimate data to represent structural changes in Asian Monsoon dynamics. Connections in the correlation network are commonly interpreted as paths of information flow. Analogous to that, variations in temperature and precipitation are conducted along atmospheric circulation patterns, transported via advection and diffusion processes. Paleoclimate archives sample the climate parameters and the reconstructed paleoclimate proxy time series thus can represent inter-regional climate variability along with local climate signals and potential measurement noise.

Our model therefore assumes an underlying flow system of simplified ASM components. We assume 3 sources, where random climate variability, noise, originates and which is then transported via advection and diffusion along the paths. The position and tranmission direction of the sources and the observation points are illustrated schematically in Fig. 2. At each point in the ASM region a local time series of climate variability is computed as the sum of the noise contributions from each of the three sources. These components are scaled with a factor that quantifies the amount of information that is transported from that source to the point of observation:

$$S_i = A(T,i)S_{\text{in1}} + B(T,i)S_{\text{in2}} + C(T,i)S_{\text{chi}} + S_{\text{noise}} \tag{1}$$

where S_i is the signal at point i, S_{in1} is the signal of the longitudinal ISM component, S_{in2} is the signal of the latitudinal ISM component and S_{chi} is the signal of the Chinese (EASM) source, S_{noise} is local observation noise and A(F,i), B(F,i) and C(F,i) are the scale factors at point i and a potential forcing F.

To compute these factors we approximate and solve the Advection-Diffusion-Equation with the

[1] Rehfeld and Molkenthin, *in prep.* is based on the analytically approximated Advection Diffusion equation, which is developed in Molkenthin and Rehfeld, *in prep.* to construct climate networks directly from simple fluid-dynamical considerations.

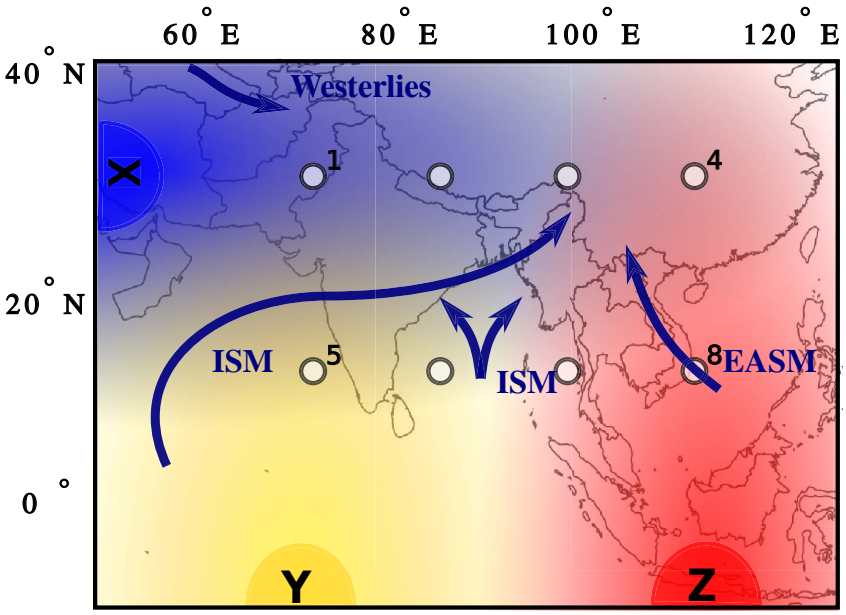


Figure 2: KIMONO: Schematic illustration of the location of sources of variability in the sampled space. The flows associated with the sources are assumed to be transmitted independently, but their local influence at the observation locations (black dots labeled) considers their relative strength at the observation points.

initial condition being a Gaussian shaped temperature front, starting from the source's position. So for example A(F,i) is the maximum height of the temperature front from the vertical Indian source, when it reaches point i.
The system we are looking at is a two dimensional boundary-less fluid of constant diffusivity χ with a stationary flow described by the velocity field $\vec{v}(\vec{x})$. Temperature transport in the system is governed by the Advection-Diffusion equation, which states how the change of temperature over time is determined by the spatial temperature change and the velocity:

$$\frac{\partial T}{\partial t} = \chi \Delta T - \nabla \cdot (\vec{v}(\vec{x})T) \ , \tag{2}$$

which is obtained by inserting the advective and diffusive flux

$$\vec{j} = \vec{j}_{diff} + \vec{j}_{adv} = -\chi \nabla T + \vec{v}T \tag{3}$$

into the sourceless continuity equation for temperature

$$\frac{\partial T}{\partial t} = -\nabla \vec{j} \ . \tag{4}$$

$T(\vec{x}, t)$ is the temperature value at position $\vec{x}$ at time t.
We use a temperature δ-peak as a tracer of the flow. It is inserted at an arbitrary point x_0 in the fluid as the initial condition, so in other words we solve the Cauchy problem of equation (2) with the initial condition

$$T(\vec{x}, 0) = \delta(\vec{x} - \vec{x_0}) \ . \tag{5}$$

For $v = const$ this can be solved analytically. We neglect the derivative of the velocity field but replace $\vec{v}$ by $\vec{v}(\vec{x})$ and thereby get an approximate solution for velocity fields with a slow spatial variation. We also assume that this velocity field depends on a given circulation forcing F.

The initial condition is a Gaussian shaped temperature front of unit height (in x- or y-direction)

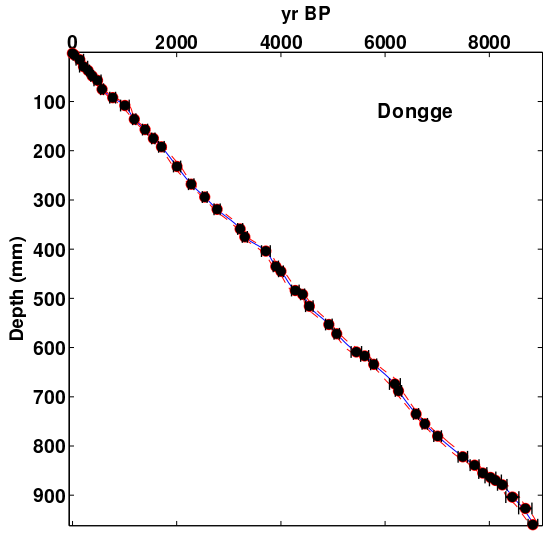
$$e^{-\frac{(x-x_0)^2}{s}}, \tag{6}$$

where s is the width and x_0 the position of the source. Local temperature is computed as a function

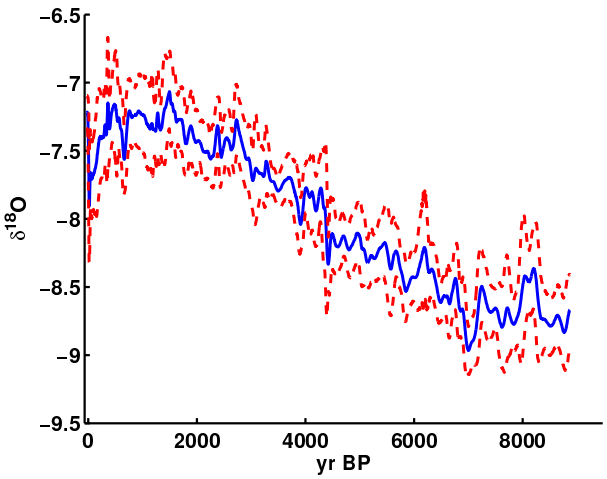
$$T(x,t) = \sqrt{\frac{s}{s+4\chi t}} e^{-\frac{(x-x_0-vt)^2}{s+4\chi t}}. \tag{7}$$

of time and space. Since this can be seen as a statistical description of how one original peak would dissipate over space we use it to define the local variance factors: $A(F,i) = T(i, t_{\max})$, with the front in y direction and $x_0 = x_{\text{in1}}$, $B(F,i) = T(i, t_{\max})$, with the front in x direction and $x_0 = x_{\text{in2}}$, $C(F,i) = T(i, t_{\max})$, with the front in x direction and $x_0 = x_{\text{chi}}$.

2 Age modeling results for paleoclimate archives

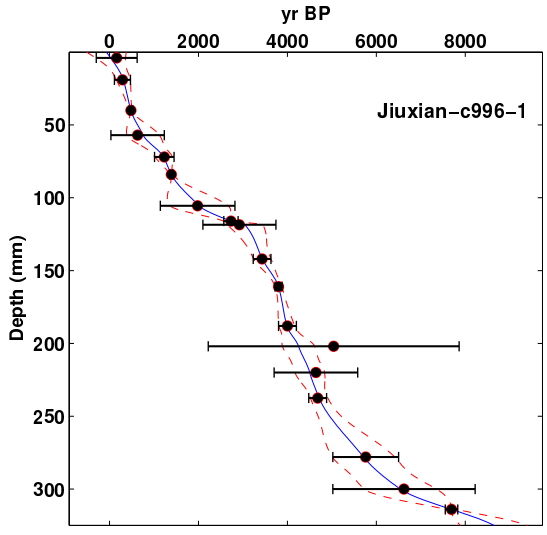


(a) Age-depth plot

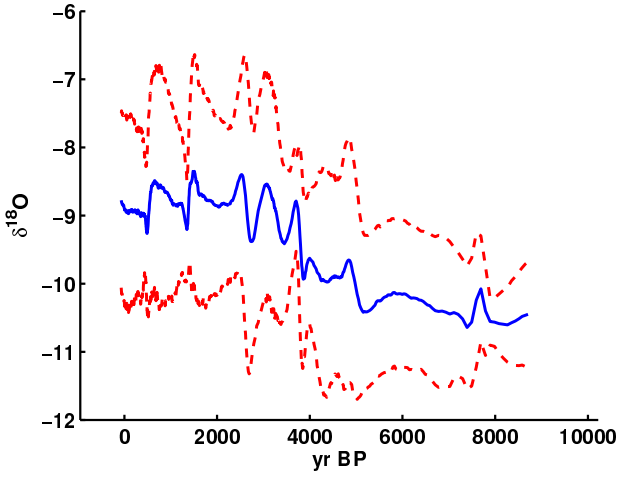


(b) Median proxy record

Figure 3: COPRA modeled depth-age relationship (a) and proxy record (b) for the Dongge cave record [175].

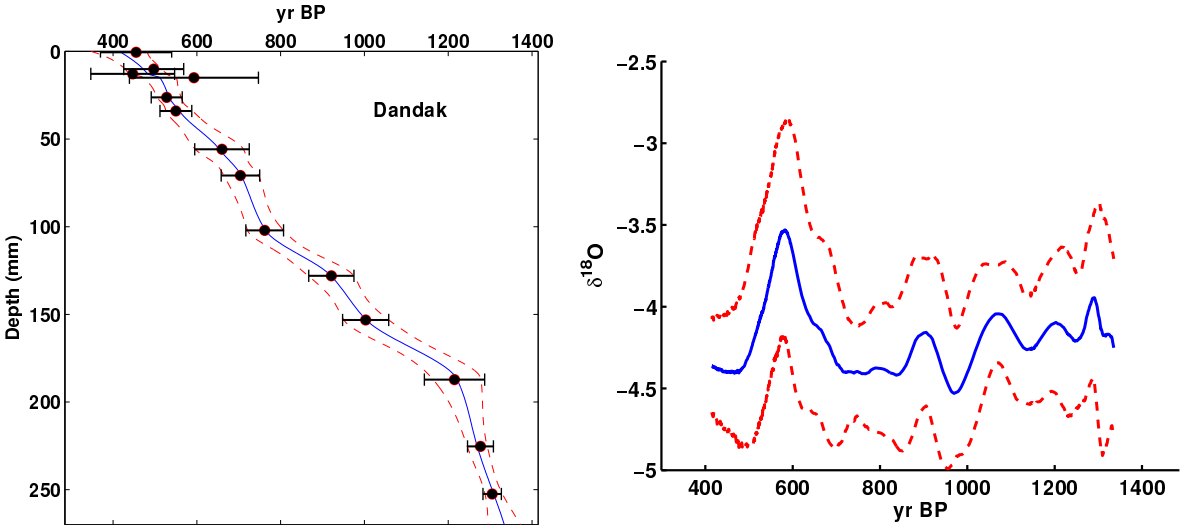


(a) Age-depth plot



(b) Median proxy record

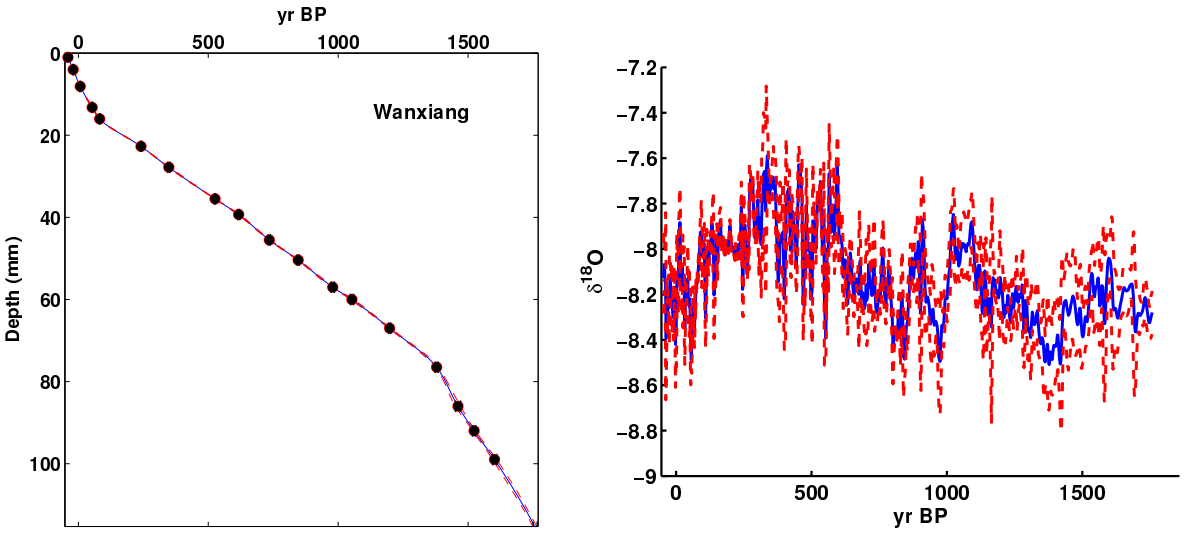
Figure 4: COPRA modeled depth-age relationship (a) and proxy record (b) for the Jiuxian-C996-1 cave record [25].



(a) Age-depth plot

(b) Median proxy record

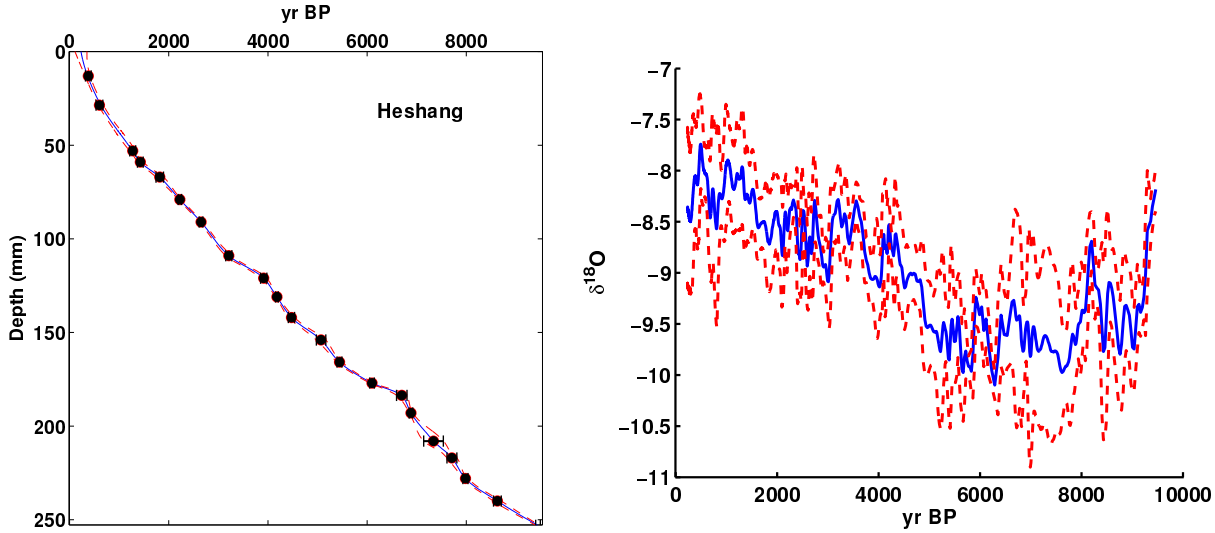
Figure 5: COPRA modeled depth-age relationship (a) and proxy record (b) for the Dandak cave record [13, 145].



(a) Age-depth plot

(b) Median proxy record

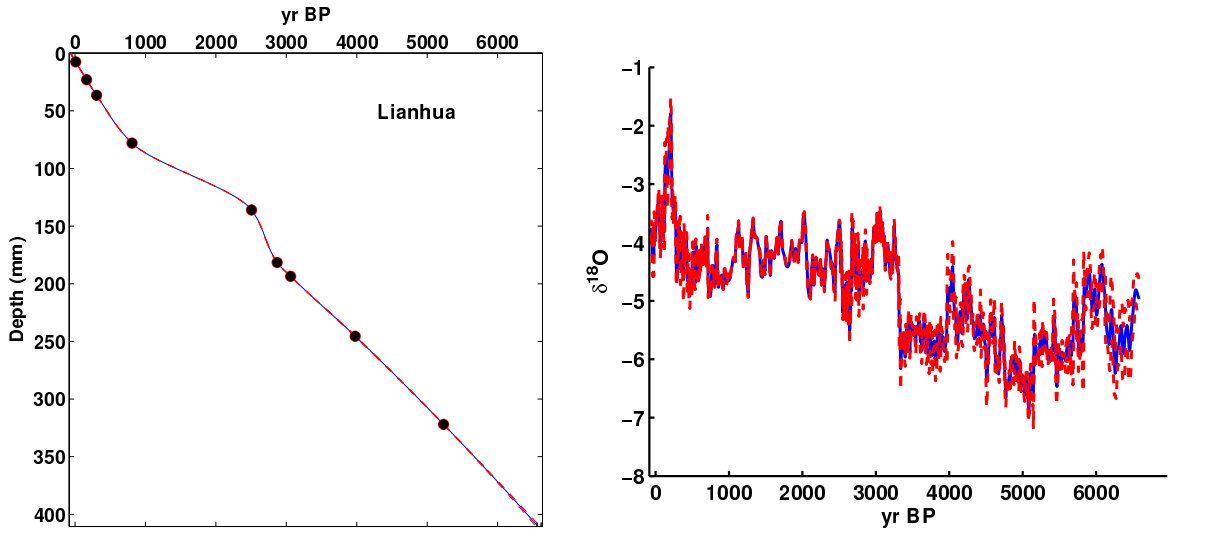
Figure 6: COPRA modeled depth-age relationship (a) and proxy record (b) for the Wanxiang cave record [190].



(a) Age-depth plot

(b) Median proxy record

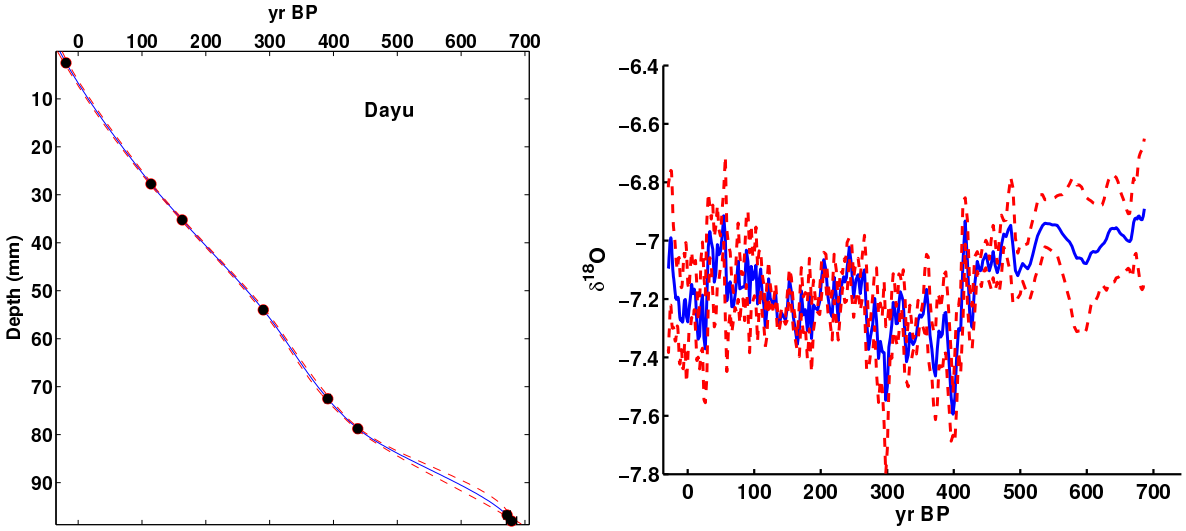
Figure 7: COPRA modeled depth-age relationship (a) and proxy record (b) for the Heshang cave record [70].



(a) Age-depth plot

(b) Median proxy record

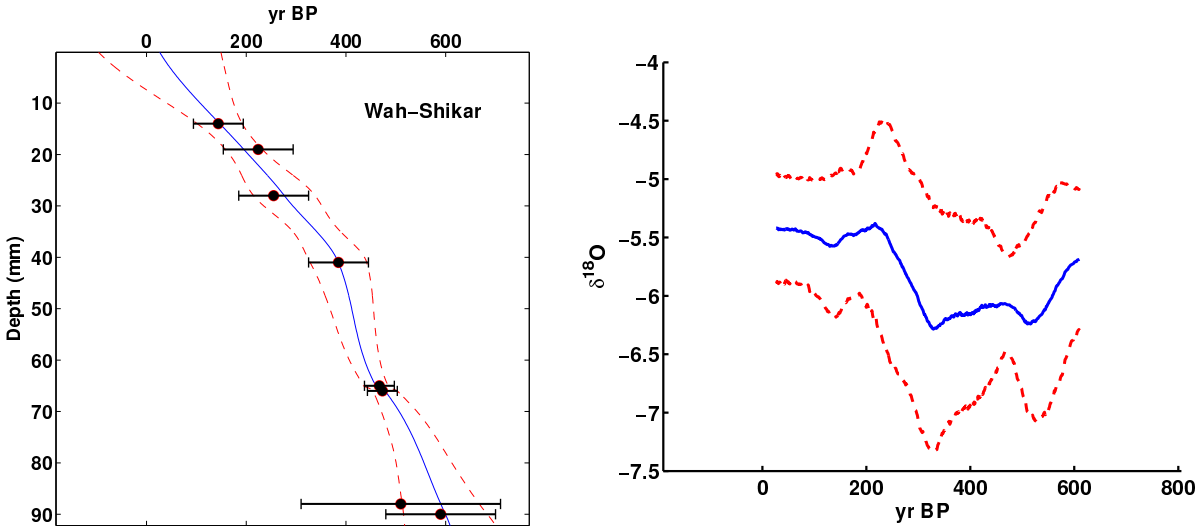
Figure 8: COPRA modeled depth-age relationship (a) and proxy record (b) for the Lianhua cave record [35].



(a) Age-depth plot

(b) Median proxy record

Figure 9: COPRA modeled depth-age relationship (a) and proxy record (b) for the Dayu cave record [157].

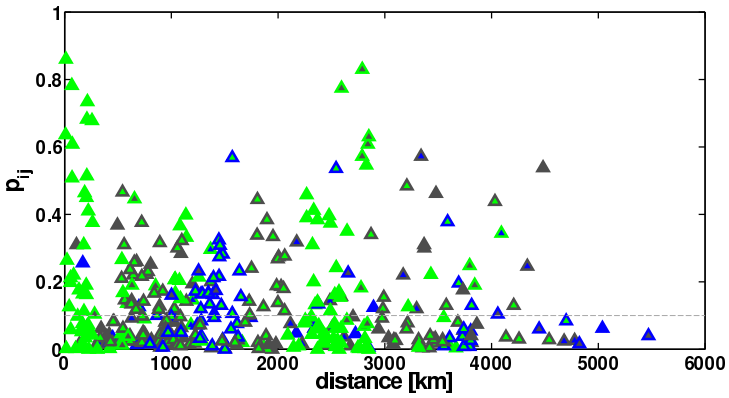


(a) Age-depth plot

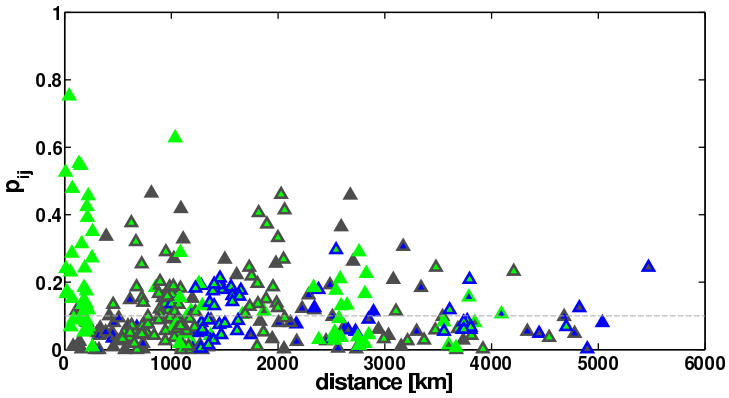
(b) Median proxy record

Figure 10: COPRA modeled depth-age relationship (a) and proxy record (b) for the Wah-Shikar cave record [146].

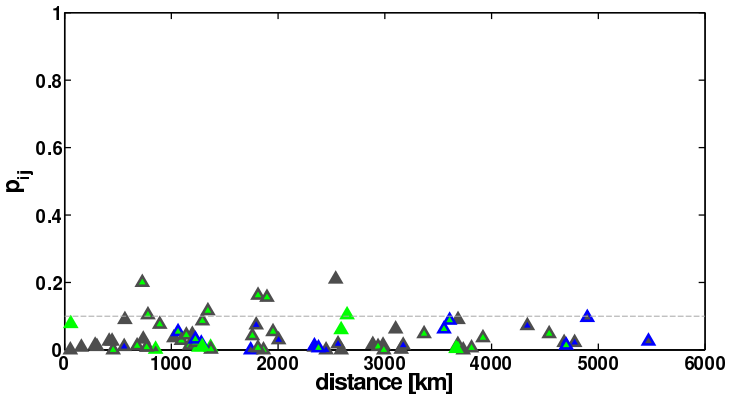
3 Link strength vs. link length in the paleoclimate network for the ASM



(a) RWP (1)



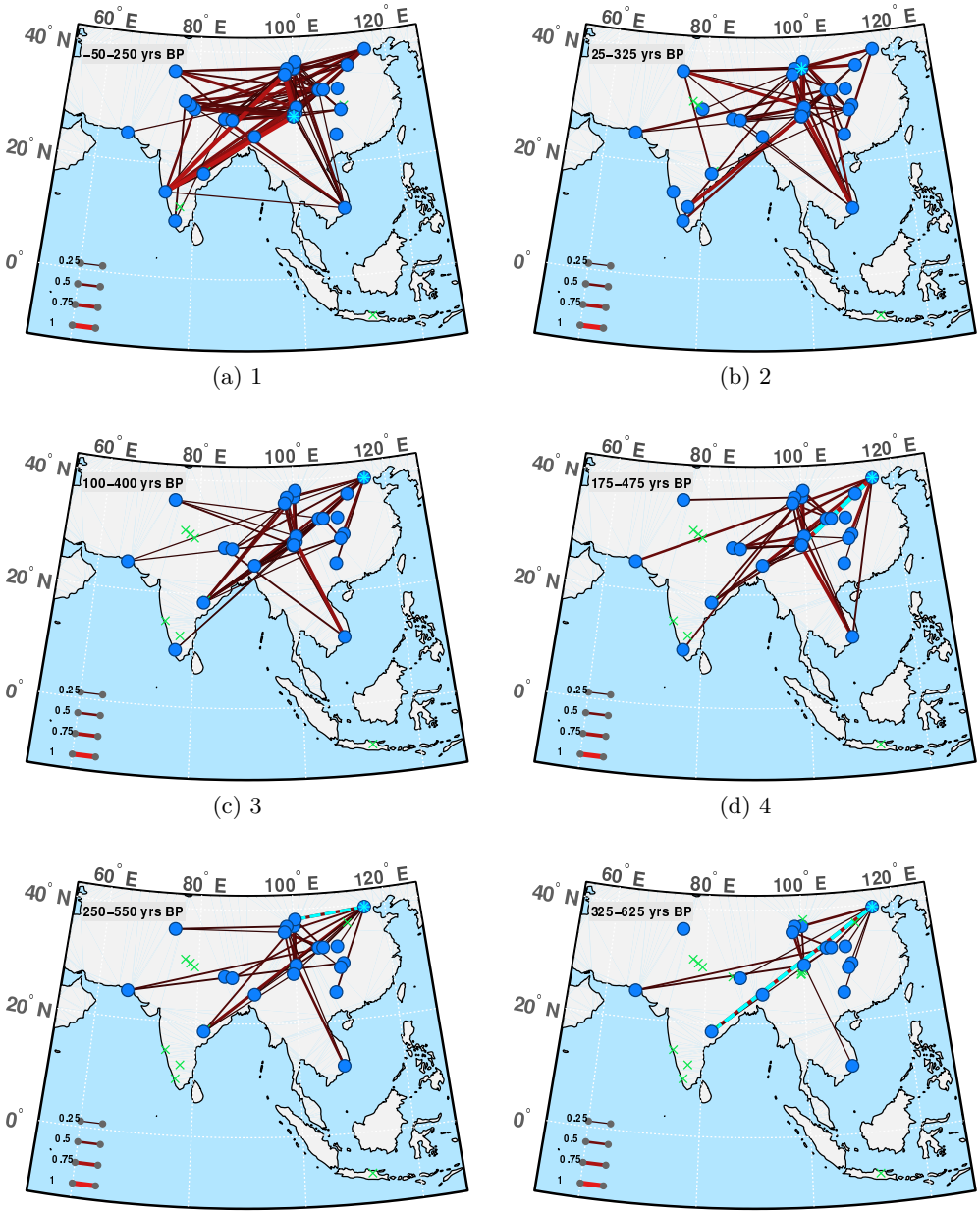
(b) LIA (3)



(c) MWP (11)

Figure 11: Link strength vs. link length for the time slices centering on the (a) RWP, (b) LIA and (c) MWP period. The two marker colors refer to the two potentially different archive types involved in a link. Uniform green rectangles denote tree-tree links, grey rectangles denote stalagmite-stalagmite links, blue parts denote that non-tree non-speleothem archive was involved.

4 ASM paleoclimate network topologies: time evolution

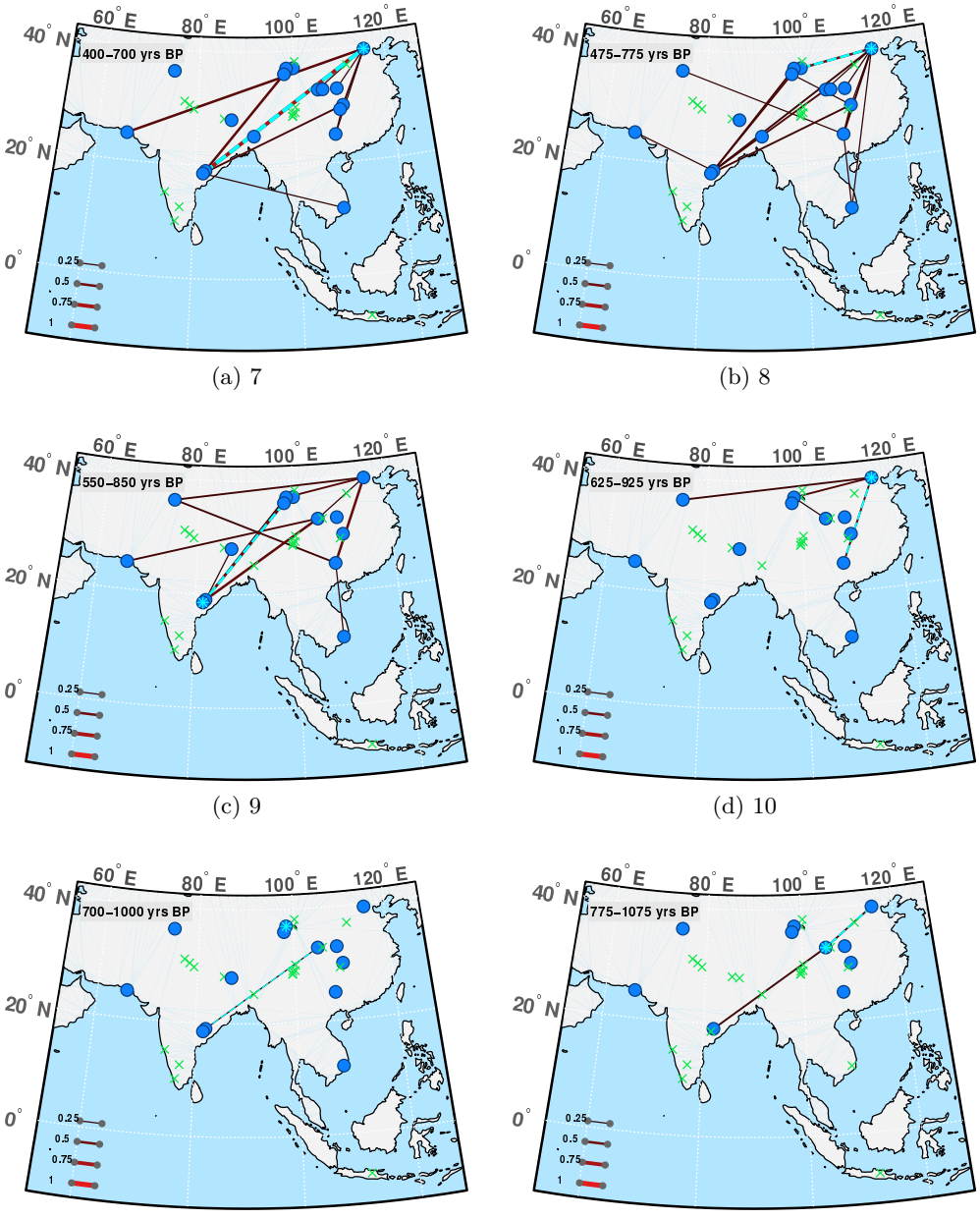


(a) 1 (b) 2

(c) 3 (d) 4

(e) 5 (f) 6

Figure 12: Paleoclimate network for the late Holocene Asian Monsoon domain and under consideration of age uncertainties: Steps 1 to 6

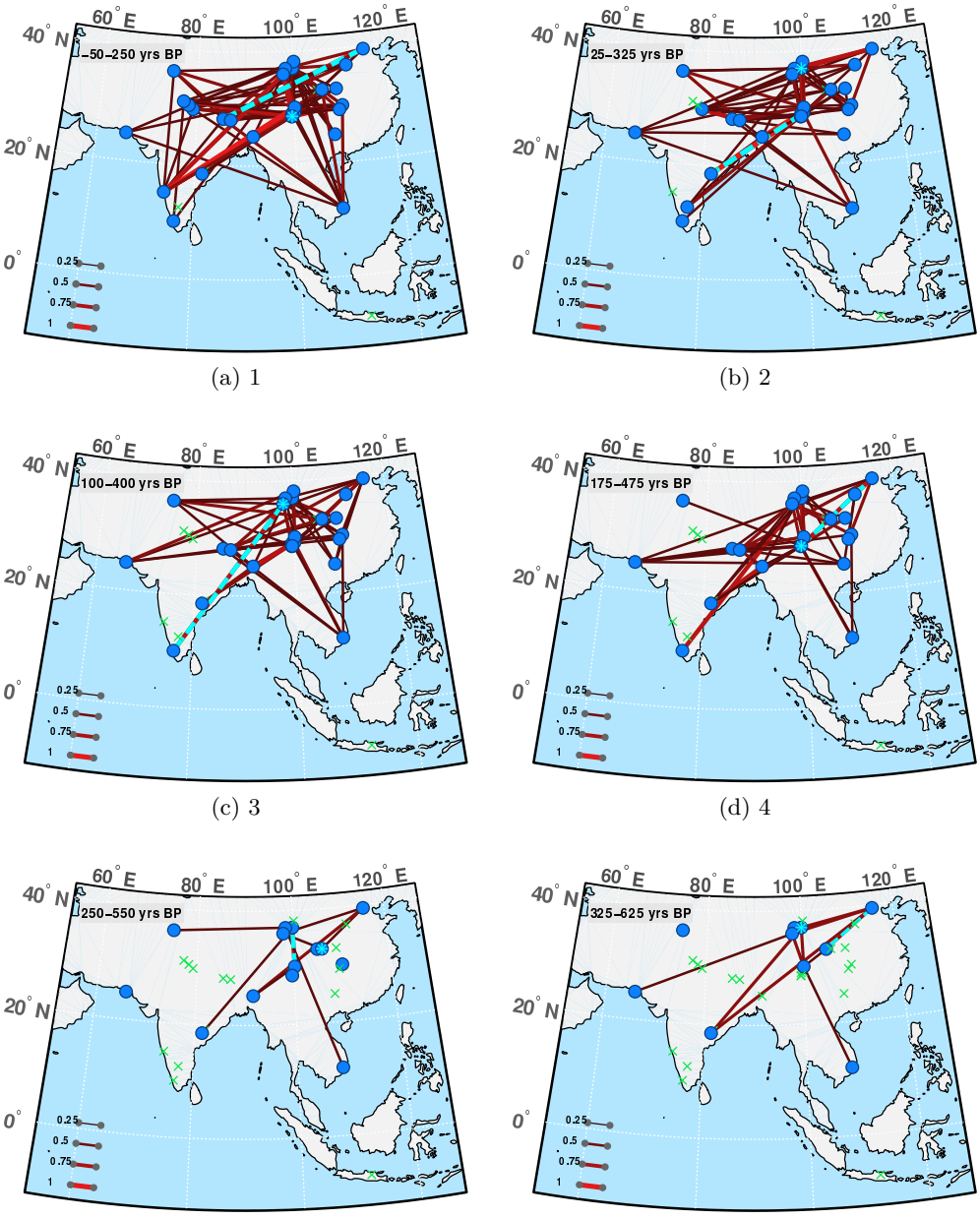


(a) 7 (b) 8

(c) 9 (d) 10

(e) 11 (f) 12

Figure 13: Paleoclimate network for the late Holocene Asian Monsoon domain and under consideration of age uncertainties: Steps 7 to 12

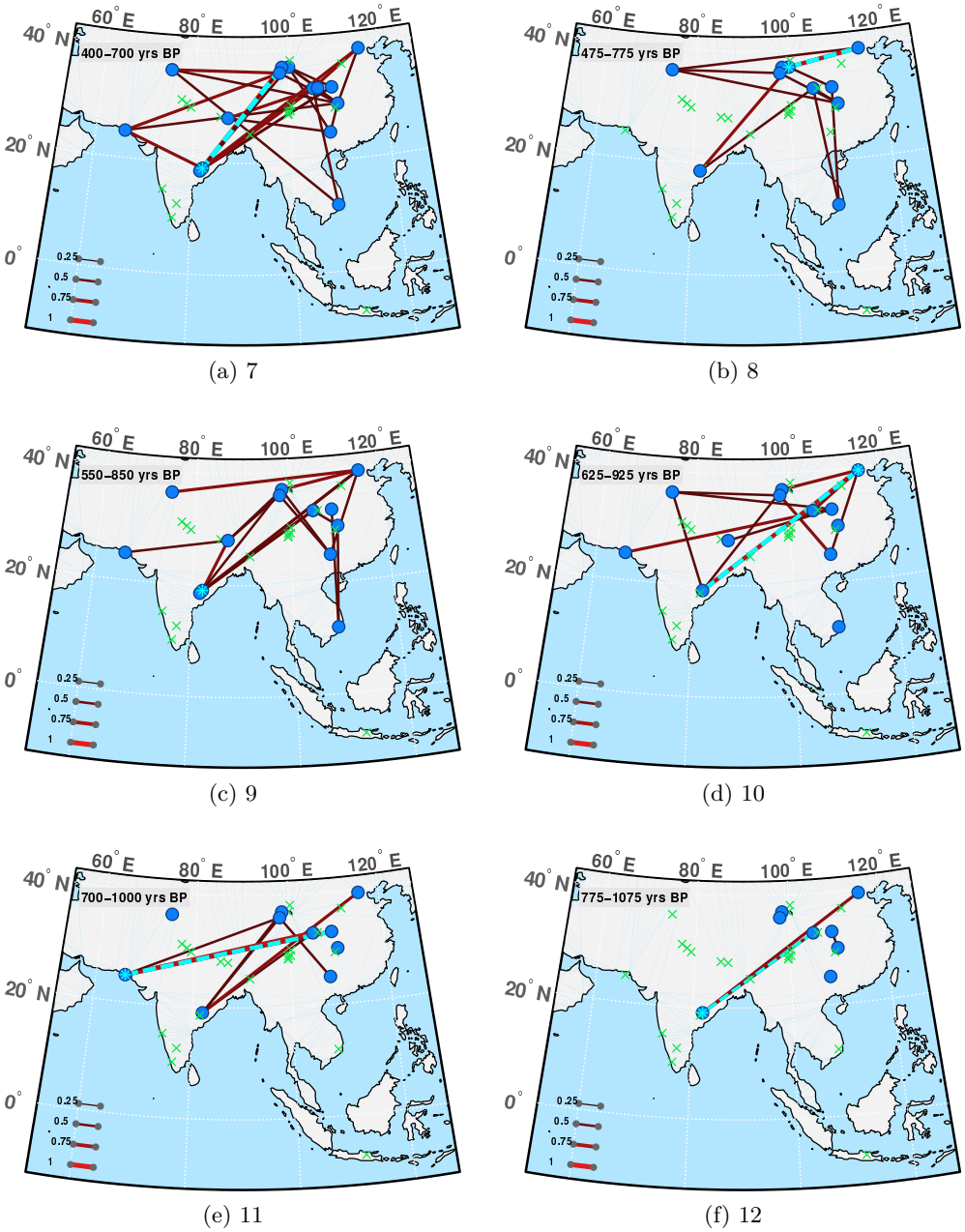


(a) 1 (b) 2

(c) 3 (d) 4

(e) 5 (f) 6

Figure 14: Paleoclimate network for the late Holocene Asian Monsoon domain without the consideration of age uncertainties: Steps 1 to 6



(a) 7 (b) 8

(c) 9 (d) 10

(e) 11 (f) 12

Figure 15: Paleoclimate network for the late Holocene Asian Monsoon domain without the consideration of age uncertainties: Steps 7 to 12

Bibliography

[1] R. Agnihotri, K. Dutta, R. Bhushan, and B. Somayajulu. Evidence for solar forcing on the Indian monsoon during the last millennium. *Earth and Planetary Science Letters*, 198 (3-4):521–527, 2002.

[2] D. J. Albers and G. Hripcsak. Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series. *Chaos, solitons, and fractals*, 45(6):853–860, 2012.

[3] K. J. Anchukaitis, M. N. Evans, A. Kaplan, E. A. Vaganov, M. K. Hughes, H. D. Grissino-Mayer, and M. A. Cane. Forward modeling of regional scale tree-ring patterns in the southeastern United States and the recent influence of summer drought. *Geophysical Research Letters*, 33(4):2–5, 2006.

[4] P. Babu and P. Stoica. Spectral analysis of nonuniformly sampled data – a review. *Digital Signal Processing*, 20(2):359–378, 2010.

[5] M. T. Bahadori and Y. Liu. Granger Causality Analysis in Irregular Time Series. In *Proceedings of the Twelfth SIAM International Conference on Data Mining*, page 12, Anaheim, CA, USA, 2012. Society for Industrial and Applied Mathematics.

[6] M. S. Baptista, I. L. Caldas, M. S. Baptista, C. S. Baptista, A. a. Ferreira, and M. V. a.P Heller. Low-dimensional dynamics in observables from complex and higher-dimensional systems. *Physica A: Statistical Mechanics and its Applications*, 287(1-2):91–99, 2000.

[7] S. Barbara, P. State, O. N. Bjoernstad, and W. Falck. Nonparametric spatial covariance functions : Estimation and testing. *Environmental and Ecological Statistics*, 8(1):53–70, 2001.

[8] M. Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.

[9] I. Batyrshin, L. Sheremetov, and J. X. Velasco-Hernandez. On axiomatic definition of time series shape association measures. In U. Villa-Vargas, L. Sheremetov, and H.-D. Haasis, editors, *Operations Research and Data Mining ORADM 2012 workshop proceedings*, pages 1–12, Mexico City, 2012. National Polytechnic Institute.

[10] L. Benedict, H. Nobach, and C. Tropea. Benchmark tests for the estimation of power spectra from LDA signals. *Proc. 9th Int. Symp. on*, 11(8):1089–1104, 1998.

[11] Y. Berezin, A. Gozolchiani, O. Guez, and S. Havlin. Stability of climate networks with time. *Scientific Reports*, 2:666, 2012.

[12] A. Berger. Milankovitch theory and climate. *Reviews of Geophysics*, 26(4):624–657, 1988.

[13] M. Berkelhammer, A. Sinha, M. Mudelsee, H. Cheng, R. L. Edwards, and K. Cannariato. Persistent multidecadal power of the Indian Summer Monsoon. *Earth and Planetary Science Letters*, 290(1-2):166–172, 2010.

[14] M. Blaauw and J. Andr. Flexible Paleoclimate Age-Depth Models Using an Autoregressive Gamma Process. *Bayesian Analysis*, 6(3):457–474, 2011.

[15] M. Blaauw, K. Bennett, and J. Christen. Random walk simulations of fossil proxy data. *The Holocene*, 20(4):645–649, 2010.

[16] H. Borgaonkar, G. Pant, and K. Rupa Kumar. Dendroclimatic reconstruction of summer precipitation at Srinagar, Kashmir, India, since the late-eighteenth century. *The Holocene*, 4(3):299–306, 1994.

[17] H. Borgaonkar, A. Sikder, S. Ram, and G. Pant. El Niño and related monsoon drought signals in 523-year-long ring width records of teak (Tectona grandis L.F.) trees from south India. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 285(1-2):74–84, 2010.

[18] R. Bos, S. de Waele, and P. Broersen. Autoregressive spectral estimation by application of the burg algorithm to irregularly sampled data. *IEEE Transactions on Instrumentation and Measurement*, 51(6):1289–1294, 2002.

[19] M. Böttcher and C. D. Dermer. Timing Signatures of the Internal-Shock Model for Blazars. *The Astrophysical Journal*, 711(1):445, 2010.

[20] S. F. Breitenbach, J. F. Adkins, H. Meyer, N. Marwan, K. K. Kumar, and G. H. Haug. Strong influence of water vapor source dynamics on stable isotopes in precipitation observed in Southern Meghalaya, NE India. *Earth and Planetary Science Letters*, 292(1-2):212–220, 2010.

[21] S. F. M. Breitenbach, K. Rehfeld, B. Goswami, J. U. L. Baldini, H. E. Ridley, D. J. Kennett, K. M. Prufer, V. V. Aquino, Y. Asmerom, V. J. Polyak, H. Cheng, J. Kurths, and N. Marwan. COnstructing Proxy Records from Age models (COPRA). *Climate of the Past*, 8(5):1765–1779, 2012.

[22] P. M. T. Broersen. Spectral Analysis of Irregularly Sampled Data with Time Series Models. *The Open Signal Processing Journal*, 1:7–14, 2008.

[23] P. M. T. Broersen and S. de Waele. The Accuracy of Time Series Analysis for Laser-Doppler Velocimetry. In *Proceedings of the 10th International Symposium on Applications of Laser Techniques to Fluid Dynamics, Lisbon, Portugal*, 2000.

[24] B. M. Buckley, K. J. Anchukaitis, D. Penny, R. Fletcher, E. R. Cook, M. Sano, L. C. Nam, A. Wichienkeeo, T. T. Minh, and T. M. Hong. Climate as a contributing factor in the demise of Angkor, Cambodia. *Proceedings of the National Academy of Sciences of the United States of America*, 107(15):6748–52, 2010.

[25] Y. Cai, L. Tan, H. Cheng, Z. An, R. L. Edwards, M. J. Kelly, X. Kong, and X. Wang. The variation of summer monsoon precipitation in central China since the last deglaciation. *Earth and Planetary Science Letters*, 291(1-4):21–31, 2010.

[26] J. Cao, J. Hu, and Y. Tao. An index for the interface between the Indian summer monsoon and the East Asian summer monsoon. *Journal of Geophysical Research*, 117(D18):1–9, 2012.

[27] S. Chakraborty, B. N. Goswami, and K. Dutta. Pacific coral oxygen isotope and the tropospheric temperature gradient over the Asian monsoon region: a tool to reconstruct past Indian summer monsoon rainfall. *Journal of Quaternary Science*, 27(3):269–278, 2012.

[28] C.-P. Chang, Y. Ding, N.-C. Lau, R. H. Johnson, B. Wang, and T. Yasunari. *The global monsoon system - Research and Forecast*. World Scientific Publishing Co, Singapore, 2 edition, 2011.

[29] C. Chatfield. *The analysis of time series: an introduction*. CRC Press, Florida, US, 6th edition, 2004.

[30] H. Cheng, A. Sinha, X. Wang, F. W. Cruz, and R. L. Edwards. The Global Paleomonsoon as seen through speleothem records from Asia and the Americas. *Climate Dynamics*, pages 1045–1062, 2012.

[31] H. Cheng, P. Z. Zhang, C. Spötl, R. L. Edwards, Y. J. Cai, D. Z. Zhang, W. C. Sang, M. Tan, and Z. S. An. The climatic cyclicity in semiarid-arid central Asia over the past 500,000 years. *Geophysical Research Letters*, 39(1):1–5, 2012.

[32] R. Clausius. Ueber die bewegende Kraft der Wärme und die Gesetze, welche sich daraus für die Wärmelehre selbst ableiten lassen. *Annalen der Physik*, 155(3):368–397, 2006.

[33] S. C. Clemens, W. L. Prell, and Y. Sun. Orbital-scale timing and mechanisms driving Late Pleistocene Indo-Asian summer monsoons: Reinterpreting cave speleothem δ 18 O. *Paleoceanography*, 25(4):PA4207, 2010.

[34] E. R. Cook, K. J. Anchukaitis, B. M. Buckley, R. D. D'Arrigo, G. C. Jacoby, and W. E. Wright. Asian monsoon failure and megadrought during the last millennium. *Science (New York, N.Y.)*, 328(5977):486–9, 2010.

[35] J. Cosford, H. Qing, B. Eglington, D. Mattey, D. Yuan, M. Zhang, and H. Cheng. East Asian monsoon variability since the Mid-Holocene recorded in a high-resolution, absolute-dated aragonite speleothem from eastern China. *Earth and Planetary Science Letters*, 275 (3-4):296–307, 2008.

[36] T. Cover and J. Thomas. *Elements of information theory*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2 edition, 2006.

[37] H. Dezhbakhsh and D. Levy. Periodic properties of interpolated time series. *Economics Letters*, 46(2):183, 1994.

[38] S. Dilmaghani, I. C. Henry, P. Soonthornnonda, E. R. Christensen, and R. C. Henry. Harmonic analysis of environmental time series with missing data or irregular sample spacing. *Environmental science & technology*, 41(20):7030–8, 2007.

[39] H. P. A. V. Dongen, E. Olofsen, J. H. VanHartevelt, E. W. Kruyt, and H. P. a. Van Dongen. Searching for Biological Rhythms : Peak Detection in the Periodogram of Unequally Spaced Data. *Journal of Biological Rhythms*, 14(6):617–620, 1999.

[40] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. *EPL (Europhysics Letters)*, 87(4):48007, 2009.

[41] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal Special Topics*, 174(1):157–179, 2009.

[42] J. F. Donges, H. C. H. Schultz, N. Marwan, Y. Zou, and J. Kurths. Investigating the topology of interacting networks. *The European Physical Journal B*, 84(4):635–651, 2011.

[43] C. Dykoski, R. Edwards, H. Cheng, D. Yuan, Y. Cai, M. Zhang, Y. Lin, J. Qing, Z. An, and J. Revenaugh. A high-resolution, absolute-dated Holocene and deglacial Asian monsoon record from Dongge Cave, China. *Earth and Planetary Science Letters*, 233(1-2):71–86, 2005.

[44] R. D'Arrigo, R. Wilson, and A. Tudhope. The impact of volcanic forcing on tropical temperatures during the past four centuries. *Nature Geoscience*, 2(1):51–56, 2008.

[45] R. A. A. Edelson and J. H. H. Krolik. The discrete correlation function - A new method for analyzing unevenly sampled variability data. *The Astrophysical Journal*, 333(October 15):646, 1988.

[46] I. Fairchild and A. Baker. *Speleothem Science: from process to past environments*. Wiley-Blackwell, 2012.

[47] J.-H. Fan, Y. Liu, B.-C. Qian, J. Tao, Z.-Q. Shen, J.-S. Zhang, Y. Huang, and J. Wang. Long-term variation time scales in OJ 287. *Research in Astronomy and Astrophysics*, 10 (11):1100, 2010.

[48] J. H. Feldhoff, R. V. Donner, J. F. Donges, N. Marwan, and J. Kurths. Geometric detection of coupling directions by means of inter-system recurrence networks. *Physics Letters A*, 376:3504–3513, 2012.

[49] D. Fleitmann. Palaeoclimatic interpretation of high-resolution oxygen isotope profiles derived from annually laminated speleothems from Southern Oman. *Quaternary Science Reviews*, 23(7-8):935–945, 2004.

[50] D. Fleitmann, S. J. Burns, A. Mangini, M. Mudelsee, J. Kramers, I. Villa, U. Neff, A. A. Al-Subbary, A. Buettner, D. Hippler, A. Matter, and A. Alsubbary. Holocene ITCZ and Indian monsoon dynamics recorded in stalagmites from Oman and Yemen (Socotra). *Quaternary Science Reviews*, 26(1-2):170–188, 2007.

[51] D. Fleitmann, H. Cheng, S. Badertscher, R. L. Edwards, M. Mudelsee, O. M. Göktürk, a. Fankhauser, R. Pickering, C. C. Raible, a. Matter, J. Kramers, and O. Tüysüz. Timing and climatic impact of Greenland interstadials recorded in stalagmites from northern Turkey. *Geophysical Research Letters*, 36(19):1–5, 2009.

[52] J. Fohlmeister. A statistical approach to construct composite climate records of dated archives. *Quaternary Geochronology*, 14:48–56, 2012.

[53] S. Gadgil. The Indian Monsoon and its Variability. *Annual Review of Earth and Planetary Sciences*, 31(1):429–467, 2003.

[54] W. Ge-Li and A. A. Tsonis. A preliminary investigation on the topology of Chinese climate networks. *Chinese Physics B*, 18(11):5091–5106, 2009.

[55] M. Ghil, M. Allen, and M. Dettinger. Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40(1):1–41, 2002.

[56] L. Giosan, P. D. Clift, M. G. Macklin, D. Q. Fuller, S. Constantinescu, J. a. Durcan, T. Stevens, G. a. T. Duller, A. R. Tabrez, K. Gangal, R. Adhikari, A. Alizai, F. Filip, S. VanLaningham, and J. P. M. Syvitski. Fluvial landscapes of the Harappan civilization. *Proceedings of the National Academy of Sciences of the United States of America*, 109(26): E1688–94, 2012.

[57] R. Govindan, D. Vyushin, A. Bunde, S. Brenner, S. Havlin, and H.-J. Schellnhuber. Global Climate Models Violate Scaling of the Observed Atmospheric Variability. *Physical Review Letters*, 89(2):028501, 2002.

[58] A. Gozolchiani, S. Havlin, and K. Yamasaki. Emergence of El Niño as an Autonomous Component in the Climate Network. *Physical Review Letters*, 107(14):1–5, 2011.

[59] A. Grinsted, J. C. Moore, and S. Jevrejeva. Reconstructing sea level from paleo and projected temperatures 200 to 2100 AD. *Climate Dynamics*, 34(4):461–472, 2009.

[60] O. Guez, A. Gozolchiani, Y. Berezin, S. Brenner, and S. Havlin. Climate network structure evolves with North Atlantic Oscillation phases. *EPL (Europhysics Letters)*, 98(3):38006, 2012.

[61] A. K. Gupta. Solar influence on the Indian summer monsoon during the Holocene. *Geophysical Research Letters*, 32(17):2–5, 2005.

[62] A. K. Gupta, D. M. Anderson, and J. T. Overpeck. Abrupt changes in the Asian southwest monsoon during the Holocene and their links to the North Atlantic Ocean. *Nature*, 421 (6921):354–7, 2003.

[63] E. Haam and P. Huybers. A test for the presence of covariance between time-uncertain series of data with application to the Dongge Cave speleothem and atmospheric radiocarbon records. *Paleoceanography*, 25(2):1–14, 2010.

[64] W. K. Harteveld, R. F. Mudde, and H. E. A. Van den Akker. Estimation of turbulence power spectra for bubbly flows from Laser Doppler Anemometry signals. *Chemical Engineering Science*, 60(22):6160–6168, 2005.

[65] U. Herzschuh. Palaeo-moisture evolution in monsoonal Central Asia during the last 50,000 years. *Quaternary Science Reviews*, 25(1-2):163–178, 2006.

[66] A. Hind, A. Moberg, and R. Sundberg. Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 2: A pseudo-proxy study addressing the amplitude of solar forcing. *Climate of the Past*, 8(4): 1355–1365, 2012.

[67] K. Hocke and N. Kämpfer. Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram. *Atmospheric Chemistry and Physics*, 9(12): 4197–4206, 2009.

[68] A. Holland and M. Aboy. A novel recursive Fourier transform for nonuniform sampled signals: application to heart rate variability spectrum estimation. *Medical & biological engineering & computing*, 47(7):697–707, 2009.

[69] Y. Hong, B. Hong, Q. Lin, Y. Zhu, Y. Shibata, M. Hirota, M. Uchida, X. Leng, H. Jiang, H. Xu, H. Wang, and L. Yi. Correlation between Indian Ocean summer monsoon and North Atlantic climate during the Holocene. *Earth and Planetary Science Letters*, 211 (3-4):371–380, 2003.

[70] C. Hu, G. Henderson, J. Huang, S. Xie, Y. Sun, and K. Johnson. Quantification of Holocene Asian monsoon rainfall from spatially separated cave records. *Earth and Planetary Science Letters*, 266(3-4):221–232, 2008.

[71] M. Hughes, T. Swetnam, and H. Diaz. *Dendroclimatology*, volume 11 of *Developments in Paleoenvironmental Research*. Springer Netherlands, Dordrecht, 2011.

[72] IPCC. The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M.Tignor, and H. Miller, editors, *Climate Change 2007*, page 996. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

[73] E. Jansen, J. Overpeck, K. Briffa, J.-C. Duplessy, F. Joos, V. Masson-Delmotte, D. Olago, B. Otto-Bliesner, W. Peltier, and S. Rahmstorf. Paleoclimate. In H. M. S. Solomon , D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, editor, *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

[74] P. D. Jones, T. J. Osborn, and K. R. Briffa. The evolution of climate over the last millennium. *Science (New York, N.Y.)*, 292(5517):662–7, 2001.

[75] H. Kantz and T. Schreiber. *Applied Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, UK, 2 edition, 2004.

[76] S. Kaspari, P. Mayewski, S. Kang, S. Sneed, S. Hou, R. Hooke, K. Kreutz, D. Introne, M. Handley, K. Maasch, D. Qin, and J. Ren. Reduction in northward incursions of the South Asian monsoon since 1400 AD inferred from a Mt. Everest ice core. *Geophysical Research Letters*, 34(16):1–6, 2007.

[77] D. J. Kennett, S. F. M. Breitenbach, V. V. Aquino, Y. Asmerom, J. Awe, J. U. L. Baldini, P. Bartlein, B. J. Culleton, C. Ebert, C. Jazwa, M. J. Macri, N. Marwan, V. Polyak, K. M. Prufer, H. E. Ridley, H. Sodemann, B. Winterhalder, and G. H. Haug. Development and Disintegration of Maya Political Systems in Response to Climate Change. *Science*, 338 (6108):788–791, 2012.

[78] D. Kondrashov and M. Ghil. Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics*, 13(2):151–159, 2006.

[79] G. S. Kong, K.-O. Kim, and S.-P. Kim. Characteristics of the East Asian summer monsoon in the South Sea of Korea during the Little Ice Age. *Quaternary International*, pages 1–9, 2012.

[80] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):1–16, 2004.

[81] T. Kreuz, D. Chicharro, R. G. Andrzejak, J. S. Haas, and H. D. I. Abarbanel. Measuring multiple spike train synchrony. *Journal of neuroscience methods*, 183(2):287–99, 2009.

[82] K. K. Kumar, K. Kamala, B. Rajagopalan, M. P. Hoerling, J. K. Eischeid, S. K. Patwardhan, G. Srinivasan, B. N. Goswami, and R. Nemani. The once and future pulse of Indian monsoonal climate. *Climate Dynamics*, 36(11-12):2159–2170, 2010.

[83] M. Küttel, J. Luterbacher, E. Zorita, E. Xoplaki, N. Riedwyl, and H. Wanner. Testing a European winter surface temperature reconstruction in a surrogate climate. *Geophysical Research Letters*, 34(7):2–7, 2007.

[84] H. Lange. Recurrence Quantification Analysis in Watershed Ecosystem Research. *International Journal of Bifurcation and Chaos*, 21(04):1113–1125, 2011.

[85] G. Lenderink and E. van Meijgaard. Increase in hourly precipitation extremes beyond expectations from temperature changes. *Nature Geoscience*, 1(8):511–514, 2008.

[86] S. Lhermitte, J. Verbesselt, W. Verstraeten, and P. Coppin. A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote Sensing of Environment*, 115(12):3129–3152, 2011.

[87] X. Liu, W. An, K. Treydte, X. Shao, S. Leavitt, S. Hou, T. Chen, W. Sun, and D. Qin. Tree-ring δ18O in southwestern China linked to variations in regional cloud cover and tropical sea surface temperature. *Chemical Geology*, 291:104–115, 2012.

[88] Z.-B. Ma, H. Cheng, M. Tan, R. L. Edwards, H.-C. Li, C.-F. You, W.-H. Duan, X. Wang, and M. J. Kelly. Timing and structure of the Younger Dryas event in northern China. *Quaternary Science Reviews*, 41:83–93, 2012.

[89] B. Maher. Holocene variability of the East Asian summer monsoon from Chinese cave records: a re-assessment. *The Holocene*, 18(6):861–866, 2008.

[90] K. Mahr. How the Changing Monsoon Is Changing India, 2012. URL `http://world.time.com/2012/07/19/how-the-changing-monsoon-is-changing-india/`.

[91] N. Malik, N. Marwan, and J. Kurths. Spatial structures and directionalities in Monsoonal precipitation over South Asia. *Nonlinear Processes in Geophysics*, 17(5):371–381, 2010.

[92] N. Malik, B. Bookhagen, N. Marwan, and J. Kurths. Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. *Climate Dynamics*, 39(3-4):971–987, 2011.

[93] N. Malik, Y. Zou, N. Marwan, and J. Kurths. Dynamical regimes and transitions in Plio-Pleistocene Asian monsoon. *EPL (Europhysics Letters)*, 97(4):40009, 2012.

[94] S. R. Managave, M. S. Sheshshayee, a. Bhattacharyya, and R. Ramesh. Intra-annual variations of teak cellulose δ18O in Kerala, India: implications to the reconstruction of past summer and winter monsoon rains. *Climate Dynamics*, 37(3-4):555–567, 2010.

[95] M. Mann, R. Bradley, and M. Hughes. Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, 392:779–788, 1998.

[96] M. E. Mann. Little Ice Age. In M. C. M. Cracken, J. S. Perry, and T. Munn, editors, *Volume 1, The Earth system: physical and chemical dimensions of global environmental change*, volume 1, pages 504–509. John Wiley & Sons, Ltd, Chichester, encycloped edition, 2002. ISBN 0471977969.

[97] M. E. Mann and S. Rutherford. Climate reconstruction using 'Pseudoproxies'. *Geophysical Research Letters*, 29(10):1501, 2002.

[98] M. E. Mann, J. D. Fuentes, and S. Rutherford. Underestimation of volcanic cooling in tree-ring-based reconstructions of hemispheric temperatures. *Nature Geoscience*, 5(3):202–205, 2012.

[99] N. Marwan. Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, 302(5-6):299–307, 2002.

[100] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, 2007.

[101] W. May. The sensitivity of the Indian summer monsoon to a global warming of 2C with respect to pre-industrial times. *Climate Dynamics*, 37(9-10):1843–1868, 2010.

[102] P. A. Mayewski, E. E. Rohling, J. Curt Stager, W. Karlén, K. A. Maasch, L. David Meeker, E. A. Meyerson, F. Gasse, S. van Kreveld, K. Holmgren, J. Lee-Thorp, G. Rosqvist, F. Rack, M. Staubwasser, R. R. Schneider, and E. J. Steig. Holocene climate variability. *Quaternary Research*, 62(3):243–255, 2004.

[103] W. Mayo. Spectrum measurements with laser velocimeters. In *Proc. Dynamic Flow Conf.*, pages 851–868.

[104] S. R. Meyers and B. B. Sageman. Detection, quantification, and significance of hiatuses in pelagic and hemipelagic strata. *Earth and Planetary Science Letters*, 224(1-2):55–72, 2004.

[105] D. Mitchell. Generating antialiased images at low sampling densities. *Computer Graphics*, 21(4):65–72, 1987.

[106] P. Moore, M. Little, and P. McSharry. Correlates of Depression in Bipolar Disorder (in press). *Proceedings of the Royal Society B: Biological Sciences*, pages 1–7, 2012.

[107] M. D. Morse and J. M. Patel. An efficient and accurate method for evaluating time series similarity. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07*, page 569, 2007.

[108] M. Mudelsee. TAUEST: a computer program for estimating persistence in unevenly spaced weather/climate time series. *Computers & Geosciences*, 28(1):69–72, 2002.

[109] M. Mudelsee. Estimating Pearson's Correlation Coefficient with Bootstrap Confidence Interval from Serially Dependent Time Series. *Mathematical Geology*, 35(6):651–665, 2003.

[110] M. Mudelsee. *Climate time series analysis*. Springer, 2010.

[111] M. Mudelsee, D. Scholz, R. Röthlisberger, D. Fleitmann, A. Mangini, and E. W. Wolff. Climate spectrum estimation in the presence of timescale errors. *Nonlinear Processes in Geophysics*, 16(1):43–56, 2009.

[112] M. Mudelsee, J. Fohlmeister, and D. Scholz. Effects of dating errors on nonparametric trend analyses of speleothem time series. *Climate of the Past Discussions*, 8(3):1973–2005, 2012.

[113] K. Nagashima and R. Tada. Teleconnection mechanism between millennial-scale Asian Monsoon dynamics and North Atlantic climate. *PAGES News*, 20(2):64–65, 2012.

[114] D. Nazareth, E. Soofi, and H. Zhao. Visualizing Attribute Interdependencies Using Mutual Information, Hierarchical Clustering, Multidimensional Scaling, and Self-organizing Maps. *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 53–53, 2007.

[115] E. Nieppola, T. Hovatta, M. Tornikoski, E. Valtaoja, M. F. Aller, and H. D. Aller. Long-Term Variability of Radio-Bright BL Lacertae Objects. *The Astronomical Journal*, 137(6): 5022, 2009.

[116] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 2010.

[117] T. J. Osborn and K. R. Briffa. The spatial extent of 20th-century warmth in the context of the past 1200 years. *Science (New York, N.Y.)*, 311(5762):841–4, 2006.

[118] G. Pant, K. R. Kumar, and H. Borgaonkar. Statistical models of climate reconstruction using tree ring data. *Proc. Indian natn. Sci. Acad.*, 54(3):354–364, 1988.

[119] A. Papana and D. Kugiumtzis. Evaluation of mutual information estimators for time series. *International Journal of Bifurcation and Chaos*, 19(12):4197–4215, 2009.

[120] B. H. Peter Hall, Nicholas I. Fisher, P. Hall, N. I. N. N. I. Fisher, and B. Hoffmann. On the Nonparametric Estimation of Covariance Functions. *The Annals of Statistics*, 22(4): 2115–2134, 1994.

[121] M. Pienaar and J. Tapson. A new approach to finding similarities between time series using Cross Wavelet Phase Variance. In F. Nicolls, editor, *Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa*, pages 7–10, Stellenbosch, South Africa, 2010.

[122] C. Ponton, L. Giosan, T. I. Eglinton, D. Q. Fuller, J. E. Johnson, P. Kumar, and T. S. Collett. Holocene aridification of India. *Geophysical Research Letters*, 39(3):1–6, 2012.

[123] R. Quian Quiroga, T. Kreuz, and P. Grassberger. Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Physical Review E*, 66(4): 041904, 2002.

[124] R. Ramesh, M. Tiwari, S. Chakraborty, S. R. Managave, M. G. Yadava, and D. K. Sinha. Retrieval of south Asian monsoon variation during the Holocene from natural climate archives. *Current Science*, 99(12):1170–1786, 2010.

[125] C. B. Ramsey. Radiocarbon Calibration and analysis of stratigraphy: The OxCal Program. *RADIOCARBON: Proceedings of the 15th International 14C Conference*, 37(2):425–430, 1995.

[126] H. Rashid and E. England. Late Glacial to Holocene Indian summer monsoon variability based upon sediment records taken from the Bay of Bengal. *Terrestrial Atmospheric and Oceanic Sciences*, 22(2):215–228, 2011.

[127] H. Rashid, B. Flower, R. Poore, and T. Quinn. A 25ka Indian Ocean monsoon variability record from the Andaman Sea. *Quaternary Science Reviews*, 26(19-21):2586–2597, 2007.

[128] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics*, 18(3): 389–404, 2011.

[129] K. Rehfeld, N. Marwan, S. F. M. Breitenbach, and J. Kurths. Late Holocene Asian Summer Monsoon dynamics from small but complex networks of palaeoclimate data. *Climate Dynamics*, 41(1):3–19, 2013.

[130] D. Rodbell, G. Seltzer, and D. Anderson. An ˜15,000-year record of El Niño-driven Alluviation in Southwestern Ecuador. *Science*, 283(5401):516–520, 1999.

[131] M. C. Romano, M. Thiel, J. Kurths, I. Z. Kiss, and J. L. Hudson. Detection of synchronization for non-phase-coherent and non-stationary data. *Europhysics Letters (EPL)*, 71 (3):466–472, 2005.

[132] M. Roulston. Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena*, 125(5759):285–294, 1999.

[133] M. Sano, R. Ramesh, M. S. Sheshshayee, and R. Sukumar. Increasing aridity over the past 223 years in the Nepal Himalaya inferred from a tree-ring d18O chronology. *The Holocene*, 2011.

[134] J. D. Scargle. Studies in Astronomical Time Series Analysis. I. Modeling Random Processes in the Time Domain. *The Astrophysical Journal Supplement Series*, 45(Jan):1–71, 1981.

[135] J. D. Scargle. Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263(December 15):835–853, 1982.

[136] J. D. Scargle. Studies in astronomical time series analysis. III - Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. *The Astrophysical Journal*, 343(August 15):874, 1989.

[137] M. Schimmel. Emphasizing Difficulties in the Detection of Rhythms with Lomb-Scargle Periodograms Non-Sinusoidal Rhythm : Bimodal Signal. *Biological Rhythm Research*, 32 (3):341–345, 2001.

[138] D. Scholz and D. L. Hoffmann. StalAge – An algorithm designed for construction of speleothem age models. *Quaternary Geochronology*, 6(3-4):369–382, 2011.

[139] D. Scholz, S. Frisia, A. Borsato, C. Spötl, J. Fohlmeister, M. Mudelsee, R. Miorandi, and A. Mangini. Holocene climate variability in north-eastern Italy: potential influence of the NAO and solar activity recorded by speleothem data. *Climate of the Past*, 8(4):1367–1383, 2012.

[140] M. Schulz and K. Stattegger. SPECTRUM: spectral analysis of unevenly spaced paleoclimatic time series. *Computers & Geosciences*, 23(9):929–945, 1997.

[141] P. R. Sheppard, P. E. Tarasov, L. J. Graumlich, K.-U. Heussner, M. Wagner, H. Sterle, and L. G. Thompson. Annual precipitation since 515 BC reconstructed from living and fossil juniper growth of northeastern Qinghai Province, China. *Climate Dynamics*, 23(7-8): 869–881, 2004.

[142] R. P. Shukla, K. C. Tripathi, A. C. Pandey, and I. Das. Prediction of Indian summer monsoon rainfall using Niño indices: A neural network approach. *Atmospheric Research*, 102(1-2):99–109, 2011.

[143] J. Singh, R. R. Yadav, and M. Wilmking. A 694-year tree-ring based rainfall reconstruction from Himachal Pradesh, India. *Climate Dynamics*, 33(7-8):1149–1158, 2009.

[144] A. Sinha, K. G. Cannariato, L. D. Stott, H.-C. Li, C.-F. You, H. Cheng, R. L. Edwards, I. B. Singh, and I. B. Sing. Variability of Southwest Indian summer monsoon precipitation during the Bølling-Allerød. *Geology*, 33(10):813–816, 2005.

[145] A. Sinha, K. G. Cannariato, L. D. Stott, H. Cheng, R. L. Edwards, M. G. Yadava, R. Ramesh, and I. B. Singh. A 900-year (600 to 1500 A.D.) record of the Indian summer monsoon precipitation from the core monsoon zone of India. *Geophysical Research Letters*, 34(16):1–5, 2007.

[146] A. Sinha, M. Berkelhammer, L. Stott, M. Mudelsee, H. Cheng, and J. Biswas. The leading mode of Indian Summer Monsoon precipitation variability during the last millennium. *Geophysical Research Letters*, 38(15):2–6, 2011.

[147] A. Sinha, L. Stott, M. Berkelhammer, H. Cheng, R. L. Edwards, B. Buckley, M. Aldenderfer, and M. Mudelsee. A global context for megadroughts in monsoon Asia during the past millennium. *Quaternary Science Reviews*, 30(1-2):47–62, 2011.

[148] J. E. Smerdon. Climate models as a test bed for climate reconstruction methods: pseudo-proxy experiments. *Wiley Interdisciplinary Reviews: Climate Change*, 3(1):63–77, 2012.

[149] M. Staubwasser. Climate change at the 4.2 ka BP termination of the Indus valley civilization and Holocene south Asian monsoon variability. *Geophysical Research Letters*, 30(8):3–6, 2003.

[150] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. An exploration of climate data using complex networks. *ACM SIGKDD Explorations Newsletter*, 12(1):25, 2010.

[151] F. Steinhilber, J. Beer, and C. Fröhlich. Total solar irradiance during the Holocene. *Geophysical Research Letters*, 36(19):1–5, 2009.

[152] F. Steinhilber, J. a. Abreu, J. Beer, and K. G. McCracken. Interplanetary magnetic field during the past 9300 years inferred from cosmogenic radionuclides. *Journal of Geophysical Research*, 115(A1):1–14, 2010.

[153] T. Stocker. *Introduction to Climate Modelling*. Advances in Geophysical and Environmental Mechanics and Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[154] P. Stoica and N. Sandgren. Spectral analysis of irregularly-sampled data: Paralleling the regularly-sampled data approaches. *Digital Signal Processing*, 16(6):712–734, 2006.

[155] Strogatz S.H. *Nonlinear dynamics and chaos with applications to physics, biology, chemistry, and engineering*. Westview Press, Cambridge, Mass., 1994.

[156] A. Svensson, K. K. Andersen, M. Bigler, H. B. Clausen, D. Dahl-Jensen, S. M. Davies, S. J. Johnsen, R. Muscheler, F. Parrenin, S. O. Rasmussen, R. Röthlisberger, I. Seierstad, J. P. Steffensen, and B. M. Vinther. A 60 000 year Greenland stratigraphic ice core chronology. *Climate of the Past*, 4(1):47–57, 2008.

[157] L. Tan, Y. Cai, H. Cheng, Z. An, and R. L. Edwards. Summer monsoon precipitation variations in central China over the past 750years derived from a high-resolution absolute-dated stalagmite. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 280(3-4):432–439, 2009.

[158] M. Tan, T. Liu, J. Hou, X. Quin, H. Zhang, and T. Li. Cyclic rapid warming on centennial-scale revealed by a 2650-year stalagmite record of warm season temperature. *Geophysical Research Letters*, 30(12), 2003.

[159] K. C. Taylor, R. B. Alley, D. a. Meese, M. K. Spencer, E. J. Brook, N. W. Dunbar, R. C. Finkel, A. J. Gow, A. V. Kurbatov, G. W. Lamorey, P. a. Mayewski, E. a. Meyerson, K. Nishiizumi, and G. a. Zielinski. Dating the Siple Dome (Antarctica) ice core by manual

and computer interpretation of annual layering. *Journal of Glaciology*, 50(170):453–461, 2004.

[160] R. Telford, E. Heegaard, and H. Birks. All age–depth models are wrong: but how badly? *Quaternary Science Reviews*, 23(1-2):1–5, 2004.

[161] L. G. Thompson, T. Yao, E. Mosley-Thompson, M. Davis, K. Henderson, and P.-N. Lin. A High-Resolution Millennial Record of the South Asian Monsoon from Himalayan Ice Cores. *Science*, 289(September):1998–2001, 2000.

[162] M. Tiwari, A. K. Singh, and R. Ramesh. High-Resolution Monsoon Records Since Last Glacial Maximum: A Comparison of Marine and Terrestrial Paleoarchives from South Asia. *Journal of Geological Research*, 2011:1–12, 2011.

[163] K. E. Trenberth. Relationships between precipitation and surface temperature. *Geophysical Research Letters*, 32(14):2–5, 2005.

[164] K. S. Treydte, G. H. Schleser, G. Helle, D. C. Frank, M. Winiger, G. H. Haug, and J. Esper. The twentieth century was the wettest period in northern Pakistan over the past millennium. *Nature*, 440(7088):1179–82, 2006.

[165] A. A. Tsonis, K. L. Swanson, and P. J. Roebber. What Do Networks Have to Do with Climate? *Bulletin of the American Meteorological Society*, 87(5):585–595, 2006.

[166] A. A. Tsonis, G. Wang, K. L. Swanson, F. A. Rodrigues, and L. D. F. Costa. Community structure and dynamics in climate networks. *Climate Dynamics*, 37(5-6):933–940, 2010.

[167] A. G. Turner and H. Annamalai. Climate change and the South Asian summer monsoon. *Nature Climate Change*, (June), 2012.

[168] R. Venkataramani and Y. Bresler. Perfect reconstruction formulas and bounds on aliasing error in sub-Nyquist nonuniform sampling of multiband signals. *IEEE Transactions on Information Theory*, 46(6):2173–2183, 2000.

[169] R. Venkataramani and Y. Bresler. Optimal sub-Nyquist nonuniform sampling and reconstruction for multiband signals. *IEEE Transactions on Signal Processing*, 49(10):2301–2313, 2001.

[170] U. Von Rad, M. Schaaf, K. Michels, H. Schulz, W. Berger, and F. Sirocko. A 5000-yr Record of Climate Change in Varved Sediments from the Oxygen Minimum Zone off Pakistan, Northeastern Arabian Sea. *Quaternary Research*, 51(1):39–53, 1999.

[171] H. von Storch, E. Zorita, J. M. Jones, Y. Dimitriev, F. González-Rouco, and S. F. B. Tett. Reconstructing past climate from noisy data. *Science (New York, N.Y.)*, 306(5696):679–82, 2004.

[172] B. Wang. *The Asian Monsoon*. Number June. Praxis Publishing Ltd, Berlin, 2006.

[173] B. Wang and LinHo. Rainy Season of the Asian–Pacific Summer Monsoon. *Journal of Climate*, 15(4):386–398, 2002.

[174] B. Wang, S. C. Clemens, and P. Liu. Contrasting the Indian and East Asian monsoons: implications on geologic timescales. *Marine Geology*, 201(1-3):5–21, 2003.

[175] P. Wang, S. Clemens, L. Beaufort, P. Braconnot, G. Ganssen, Z. Jian, P. Kershaw, and M. Sarnthein. Evolution and variability of the Asian monsoon system: state of the art and outstanding issues. *Quaternary Science Reviews*, 24(5-6):595–629, 2005.

[176] Y. Wang, H. Cheng, R. L. Edwards, Y. He, X. Kong, Z. An, J. Wu, M. J. Kelly, C. a. Dykoski, and X. Li. The Holocene Asian monsoon: links to solar changes and North Atlantic climate. *Science (New York, N.Y.)*, 308(5723):854–7, 2005.

[177] Y. Wang, X. Liu, and U. Herzschuh. Asynchronous evolution of the Indian and East Asian Summer Monsoon indicated by Holocene moisture patterns in monsoonal central Asia. *Earth-Science Reviews*, 103(3-4):135–153, 2010.

[178] Y. J. Wang, H. Cheng, R. L. Edwards, Z. S. An, J. Y. Wu, C. C. Shen, and J. a. Dorale. A high-resolution absolute-dated late Pleistocene Monsoon record from Hulu Cave, China. *Science (New York, N.Y.)*, 294(5550):2345–8, 2001.

[179] J. Webster, G. Brook, and L. Railsback. Stalagmite evidence from Belize indicating significant droughts at the time of Preclassic Abandonment, the Maya Hiatus, and the Classic Maya collapse. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 250:1 – 17, 2007.

[180] G. Wefer, W. Berger, J. Bijma, and G. Fischer. Clues to Ocean History: a Brief Overview of Proxies. In G. Fischer and G. Wefer, editors, *Use of Proxies in Paleoceanography*, chapter 1, pages 1–68. Springer, Heidelberg, 1999. ISBN 3-540-66340-1.

[181] Z. Wu, B. Wang, J. Li, and F. Jin. An empirical seasonal prediction model of the East Asian summer monsoon using ENSO and NAO. *Journal of Geophysical Research*, 114:1–13, 2009.

[182] M. Yadava, R. Ramesh, and G. Pant. Past monsoon rainfall variations in peninsular India recorded in a 331-year-old speleothem. *The Holocene*, 14(4):517–524, 2004.

[183] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate Networks around the Globe are Significantly Affected by El Niño. *Physical Review Letters*, 100(22):228501, 2008.

[184] B. Yang. General characteristics of temperature variation in China during the last two millennia. *Geophysical Research Letters*, 29(9):1324, 2002.

[185] X. Yang, T. Yao, W. Yang, W. Yu, and D. Qu. Co-existence of temperature and amount effects on precipitation δ18O in the Asian monsoon region. *Geophysical Research Letters*, 38(21):L21809, 2011.

[186] L. Yi, H. Yu, J. Ge, Z. Lai, X. Xu, L. Qin, and S. Peng. Reconstructions of annual summer precipitation and temperature in north-central China since 1470AD based on drought/flood index and tree-ring records. *Climatic Change*, 110(1-2):469–498, 2011.

[187] D. Yihui and J. C. L. Chan. The East Asian summer monsoon: an overview. *Meteorology and Atmospheric Physics*, 89(1-4):117–142, 2005.

[188] B.-K. Zhang, B.-Z. Dai, L. Zhang, and Z. Cao. Multi-band optical variability of BL Lac object OQ 530. *Research in Astronomy and Astrophysics*, 10(7):653, 2010.

[189] J. Zhang, F. Chen, J. A. Holmes, H. Li, X. Guo, J. Wang, S. Li, Y. Lü, Y. Zhao, and M. Qiang. Holocene monsoon climate documented by oxygen and carbon isotopes from lake sediments and peat bogs in China: a review and synthesis. *Quaternary Science Reviews*, 30(15-16):1973–1987, 2011.

[190] P. Zhang, H. Cheng, R. L. Edwards, F. Chen, Y. Wang, X. Yang, J. J. J. J. Liu, M. Tan, X. Wang, C. An, Z. Dai, J. Zhou, D. Zhang, J. Jia, L. Jin, and K. R. Johnson. A test of climate, sun, and culture relationships from an 1810-year Chinese cave record. *Science*, 322(5903):940–2, 2008.

[191] T. Zhou, B. Li, W. Man, L. Zhang, and J. Zhang. A comparison of the Medieval Warm Period, Little Ice Age and 20th century warming simulated by the FGOALS climate system model. *Chinese Science Bulletin*, 56(28-29):3028–3041, 2011.

[192] G. A. Zielinski. Use of paleo-records in determining variability within the volcanism–climate system. *Quaternary Science Reviews*, 19(1-5):417–438, 2000.

[193] R. Zu. Wulan tree ring data (ITRDB CHIN001), 1986. URL `http://www.ncdc.noaa.gov/paleo/metadata/noaa-tree-5408.html`.

[194] R. Zu. Quilan tree ring data (ITRDB CHIN003), 1986. URL `http://www.ncdc.noaa.gov/paleo/metadata/noaa-tree-5407.html`.

List of Figures

List of Tables

List of abbreviations

ACF Autocorrelation function

AR(1) Autoregressive process (first order)

ASM Asian summer monsoon

BP Before present (geosci.: time before 1950 AD)

CCF Crosscorrelation function, see also: *XCF*

COPRA COnstructing Proxy Record Agemodels [21]

EASM East Asian summer monsoon

ENSO El-Niño/Southern Oscillation

ES(F) Event synchronization (function)

FFT Fast Fourier Transform

GCM Global climate model

ISM Indian summer monsoon

ITCZ Intertropical Convergence Zone

KIMONO Toy model (Ch. 4)

LIA Little Ice Age

LS(FT) Lomb-Scargle (Fourier tranform)

MI Mutual information, see also: *XCF*

MWP Medieval Warm Period

PAN Paleoclimate network

RMSE Root Mean Square Error

RWP Recent Warm Period

XCF Cross-Correlation function, prefix dependent on estimation routine:

- iXCF - interpolation-based,
- gXCF - Gaussian-kernel based.

List of publications

2012

- K. Rehfeld, N. Marwan, S. F. M. Breitenbach, J. Kurths, and C. Dynamics. Late Holocene Asian Summer Monsoon dynamics from small but complex networks of palaeoclimate data (online first), *Climate Dynamics*, Sept. 2012.
- S. F. M. Breitenbach, K. Rehfeld, B. Goswami, J. U. L. Baldini, H. E. Ridley, D. J. Kennett, K. M. Prufer, V. V. Aquino, Y. Asmerom, V. J. Polyak, H. Cheng, J. Kurths, and N. Marwan. COnstructing Proxy Records from Age models (COPRA), *Climate of the Past*, 8(5):1765–1779, Oct. 2012.
- Y. Lyatskaya, K. Rehfeld, H. Caglar, D. V. Kadam, L. M. Chin, J. H. Killoran, and A. Allen. Comparison of two techniques for target motion evaluation based on 4DCT images, *International Journal of Biomedical Engineering and Technology (IJBET)* 8(2/3):117–137, 2012.

2011

- K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths. Comparison of correlation analysis techniques for irregularly sampled time series, *Nonlinear Processes in Geophysics*, 18(3):389–404, June 2011.
- J. F. Donges, R. V. Donner, K. Rehfeld, N. Marwan, M. H. Trauth, and J. Kurths. Identification of dynamical transitions in marine palaeoclimate records by recurrence network analysis, *Nonlinear Processes in Geophysics* 18(5):545–562, Sept. 2011.

Acknowledgements

First of all I would like to thank Prof. Kurths for his trust and exceptional support, from the beginning to the end. Also, facing, and recognizing, the intricate traps paleoclimate reconstruction holds would have been impossible without Norbert Marwan and Jobst Heitzig. Without them I might have gone lost in the jungle of methods and estimators, statistics and significance, and it would have been a lot less fun. A big THANKS goes to Sebastian Breitenbach for inspiring and motivating me to dig into the Asian monsoon and the messy datasets, and for taking precious time to check the my physicist thoughts on the monsoon, making sure that my writing stays accessible for geoscientists. Nora Molkenthin is truly fearless – especially and also when it comes to nasty equations. Our crazy afternoon idea – using her approximation for the advection-diffusion-equation to model spatial climate variability for the Asian monsoon – really developed well on our shared couch. Franziska Lechleitner assisted by running COPRA for the dated archives in the Asian monsoon database – a task that saved me a lot of time when I really could use it. Cornelia Strube helped by porting the COPRA software from Matlab to Octave, and in combining it with the benchmark tests – your questions and comments made things much easier in the end. I'm greatful to Bedartha Goswami for being a really good discussion partner when it comes to anything from paleoclimatology to nonlinear dynamics, and enthusiasm kept us going until it was possible to find solutions. Hemant Borgaonkar, Madhusan Yadava, Prof. R. Ramesh, Yanjun Cai and Max Berkelhammer provided me with datasets, which were especially crucial as there are so few of them for the Asian monsoon. I'm very grateful also to Gerd Helle and Mandy Freud for very helpful information on tree-ring data and dendroclimatology in general. Jeff Scargle is acknowledged for providing me with very valuable and helpful comments on (kernel-based) correlation and spectral estimation – I admit I was relieved to find he took the "defeat" of the Lomb-Scargle Periodogram technique for cross-correlation analysis well – and amused to find that, in astronomy, kernel-based estimators have been in use for quite a while! Big thanks also go to my proof readers: Matze, Claudine, Bedartha, Peter, Jonathan, Sebastian, Jobst, Jakob, Nora, Norbert and Carsten – it was amazing to see how much the text profited. Then, Reik, Veronika, Liuba, Alex, Silvana, Nishant, Micha, Thomas, Aljoscha, Hannes, Gabi and Heike – I'm glad I work with you guys! I'm very glad I could take part in the Graduate Program Adlershof, which supports female PhD students in the sciences at Humboldt Universität zu Berlin. In this context, I thank Prof. Maik Thomas for the interesting, instructive and helpful discussions in our mentoring meetings.

My family made me who I am and I am endlessly grateful for that! Finally, my love goes out to you, Matze, I see wonderous journeys and adventures ahead – and it takes more than 42.195 kilometers to tire us, I know.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 7.01.2013 Kira Rehfeld